Application of Cognitive, Skill-Based, and Affective Theories of Learning Outcomes to New Methods of Training Evaluation

Kurt Kraiger, J. Kevin Ford, and Eduardo Salas

Although training evaluation is recognized as an important component of the instructional design model, there are no theoretically based models of training evaluation. This article attempts to move toward such a model by developing a classification scheme for evaluating learning outcomes. Learning constructs are derived from a variety of research domains, such as cognitive, social, and instructional psychology and human factors. Drawing from this research, we propose cognitive, skill-based, and affective learning outcomes (relevant to training) and recommend potential evaluation measures. The learning outcomes and associated evaluation measures are organized into a classification scheme. Requirements for providing construct-oriented evidence of validity for the scheme are also discussed.

Training evaluation is the systematic collection of data regarding the success of training programs (I. L. Goldstein, 1986). Constructive evaluation occurs when specified outcome measures are conceptually related to intended learning objectives. Evaluation is conducted to answer either of two questions: whether training objectives were achieved (learning issues), and whether accomplishment of those objectives results in enhanced performance on the job (transfer issues).

Although both questions are important, Campbell (1988) stressed that the most fundamental issue of evaluation is whether trainees have learned the material covered in training. Shuell (1986) also argued that evaluation should be conducted first to determine whether the intended outcomes of training have been achieved. Unfortunately, there are no conceptual models to guide researchers in decisions about how to evaluate learning. Therefore, this article focuses on issues related to specification and measurement of learning. Our objective is to derive a conceptually based scheme for evaluating learning outcomes; this classification scheme would then serve as a starting point for the development of a true training evaluation model.

The assumptions underlying our approach are that learning outcomes are multidimensional and that progress in the training field requires taking a construct-oriented approach to learning. By multidimensional, we mean that learning may be evident from changes in cognitive, affective, or skill capacities. This perspective contrasts approaches taken in other assessment models, which have either ignored learning outcomes or have treated learning as a unidimensional construct. The need for a construct-oriented approach to learning is consistent with the views of various reviewers of the training literature who have continually bemoaned the absence of theory in training research (Campbell, 1971; I. L. Goldstein, 1980; Latham, 1988; Wexley, 1984). Whereas theoretical progress has been made in other training subsystems (Cannon-Bowers, Tannenbaum, Salas, & Converse, 1991; Tannenbaum & Yukl, 1992), less progress has been made in the area of evaluation.

The absence of a conceptual basis for evaluating learning is characteristic of prior models. Historically, the most popular evaluation model, proposed by Kirkpatrick (1976, 1987), identifies four levels of evaluation: trainee reactions, learning, behavior, and organizational results. Within this model, learning is measured by examining the extent to which trainees have acquired relevant principles, facts, or skills and could be assessed using traditional multiple-choice tests. Learning was conceptualized both as a causal result of positive reactions to training and as a causal determinant of changes in trainee behavior (Alliger & Janak, 1989).

Although there are a number of conceptual flaws in the model (Alliger & Janak, 1989; Clement, 1982; Snyder, Raben, & Farr, 1980), in the present context, its greatest shortcomings are a lack of clarity regarding what specific changes may be expected as a function of trainee learning and the difficulty in identifying what assessment techniques are appropriate given
those expectations. For example, it is not clear whether Kirkpatrick (1976, 1987) thought of learning skills and learning facts as synonymous and whether the same assessment tools are appropriate for each. As Kirkpatrick's recommendations continue to represent state-of-the-art training evaluation, there has been insufficient research as to what constitutes learning and how learning outcomes should be assessed.

More recently, other researchers have embedded learning indices into conceptual models of training effectiveness. The phrases training evaluation and training effectiveness have often been used interchangeably (e.g., Ostroff, 1991), yet each address very different research questions. Training evaluation refers to a system for measuring whether trainees have achieved learning outcomes. It is concerned with issues of measurement and design, the accomplishment of learning objectives, and the attainment of requisite knowledge and skills. In contrast, training effectiveness models seek to explicate why training did or did not achieve its intended outcomes. This objective is accomplished by identifying and measuring the effects of individual, organizational, and training-related factors on training outcomes such as learning and transfer of training (Tannenbaum, Mathieu, Salas, & Cannon-Bowers, 1991). Issues of transfer of training and training effectiveness are necessarily broader than issues of training evaluation. However, it is critical to recognize that these models require substantive criterion variables provided by good evaluation theories. It is equally important to note that conceptually sound measures of learning have been absent from these models.

An examination of training effectiveness models illustrates the central role of learning. For example, Baldwin and Ford (1988) discussed trainee, training design, and organizational characteristics that may affect the transfer of training. They included learning and retention as mediating variables between these characteristics and the generalization and maintenance of trained behaviors on the job. I. L. Goldstein (1991; Rouillier & Goldstein, 1991) derived a transfer climate model that identified and tested specific situational cues and consequences in the organizational environment that affect transfer. The amount of learning obtained by the trainee was recognized as an important precursor to transfer. Whereas the work of these authors highlights the importance of learning and transfer issues, we believe additional benefits can be derived from multidimensional perspectives on learning and learning outcomes.

Finally, Noe (1986) and Tannenbaum et al. (1991) derived broader models of training effectiveness that identified situational and individual factors that affected both learning and posttraining outcomes; learning during training was again proposed as a mediating variable between situational factors and desired organizational outcomes. However, even in these models, learning is treated as a unidimensional construct and different learning outcomes are not defined.

Thus, despite the critical role of trainee learning in the training process, previous models of training evaluation (e.g., Kirkpatrick, 1976, 1987) and transfer of training or training effectiveness (e.g., I. L. Goldstein, 1991; Tannenbaum et al., 1991) have defined it in a simplistic and unidimensional way. To date, there has been little attention given to conceptualizing and measuring these learning outcomes.

Toward a Classification Scheme of Learning Outcomes

Generally, the training field has envisioned learning outcomes solely as changes in verbal knowledge or behavioral capacities. These learning outcomes are unnecessarily restrictive and out of step with modern learning theories. In other disciplines, more complex and extensive taxonomies of learning outcomes highlight the true multifaceted nature of learning (Bloom, 1956; Gagne, 1984; Krathwohl, Bloom, & Masia, 1964). Decades ago, Bloom (1956) proposed that cognitive outcomes beyond recall or recognition of verbal knowledge are legitimate learning outcomes and proposed a taxonomy of cognitively based learning outcomes. Krathwohl et al. (1964) expanded this taxonomy further to include affect-oriented objectives such as awareness and appreciation. Gagne (1984) was also critical of limiting instructional objectives to the behavioral domain and reinforced the need to examine various cognitive, skill-oriented, and attitudinal learning outcomes.

To advance the science and practice of training evaluation, it is necessary to move toward a conceptually based classification scheme of learning based on such a multidimensional perspective. Drawing from Bloom's (1956) and Gagne's (1984) taxonomies, we propose three categories of learning outcomes: cognitive, skill-based, and affective.

Figure 1 presents an overview of the three learning outcomes and the constructs that are the focus of this article. In line with Gagne (1984), the learning constructs most relevant for the cognitive category include verbal knowledge, knowledge organization, and cognitive strategies. Skill-based outcomes include skill compilation and automaticity. Finally, attitudinal outcomes (attitude, object, and strength) and motivational outcomes such as disposition, self-efficacy, and goal setting are proposed as key affective learning outcomes.

For each of the three categories, we will review relevant theory or research from a wide variety of psychological domains and identify constructs that are indicators of each learning outcome. Measurement issues relevant for each learning construct are provided and methods for training evaluation

\[ \text{LEARNING} \]

- Cognitive Outcomes
  - Verbal knowledge
  - Knowledge organization
  - Cognitive strategies

- Skill-Based Outcomes
  - Compilation
  - Proceduralization
  - Composition
  - Automaticity

- Affective Outcomes
  - Attitudinal
  - Motivational
  - Disposition
  - Self-efficacy
  - Goal setting

\[ \text{Figure 1. A preliminary classification scheme of learning outcomes.} \]
proposed. A summary classification scheme is provided in the discussion section, following the presentation of all outcomes.

**Cognitive Learning Outcomes**

*Cognition* refers to a class of variables related to the quantity and type of knowledge and the relationships among knowledge elements. An important goal of cognitive science is to generate theoretical systems or models that specify how people function (Lord & Maher, 1991). With respect to evaluation, a cognitive perspective focuses not only on static states of trainee knowledge, but on the dynamic processes of knowledge acquisition, organization, and application.

Traditionally, knowledge acquisition in the training domain has been assessed by achievement tests administered at the end of training. Trainees may be presented with a series of questions in either multiple-choice or true–false format and are required to indicate whether each stimulus exists in memory. For example, matching the term *active listening* to the clause "use of body language, empathy, and paraphrasing to discern content and meaning" would reflect knowledge of a concept covered in a supervisory skills training course. Such formats are best suited for testing the retention of verbal or declarative knowledge (Gagne, 1977).

In contrast, research in other psychological domains has highlighted the complex and dynamic nature of the knowledge acquisition process. Although the acquisition of verbal knowledge serves as a foundation for cognitive skill development (Anderson, 1982), measures of verbal knowledge may be unable to discriminate among learners at higher levels of cognitive development. Thus, we have adapted Gagné's (1984) three categories of cognitive learning objectives to suggest three general categories of cognitively based evaluation measures: verbal knowledge, knowledge organization, and cognitive strategies. Whereas all three outcomes may be useful for evaluating trainees at any level of development, generally the three are ordered chronologically with respect to anticipated changes in trainees. That is, measures of verbal knowledge would be the most sensitive for trainees during the initial stages of skill acquisition, and strategy-based measures would be more useful for advanced trainees.

In the following sections, we will discuss measurement issues relevant to each cognitively based measure. We have purposely presented less information on verbal knowledge measures, given that these are fairly common in the training field. Instead, we have focused our presentation on theory and applications related to the other two categories.

**Verbal Knowledge**

Encoded knowledge exists in different forms, including declarative knowledge (information about what), procedural knowledge (information about how; Anderson, 1982), and strategic (Greeno, 1980) or tacit knowledge (information about which, when, and why; Wagner, 1987). Most theories of cognitive skill development agree that the acquisition of declarative knowledge must precede higher order development (Ackerman, 1987; Anderson, 1982; Fitts & Posner, 1967). Accordingly, specifying and measuring trainees' retention of declarative knowledge is most appropriate in the initial stages of training.

However, it should be noted that developing a foundation of verbally based, task-relevant knowledge is a necessary but not sufficient condition for higher order skill development. Ackerman (1986, 1987) argued convincingly that the latent abilities influencing task performance differ according to the developmental stage. Specifically, general intelligence factors seem to be the most critical for determining performance on novel tasks; trainees competent at inferring relations or memorizing information will be more successful early in training. However, with continued practice, task behaviors become internalized, and performance levels are influenced as much by psychomotor differences (e.g., speed of encoding or responding) as they are by general intellectual capabilities (Ackerman, 1987). Clearly, the danger is that whereas typical paper-and-pencil tests may be the most appropriate measure for assessing trainees' initial learning, the conceptual relationship between learning and achievement may be confounded by trainees' general intelligence. That is, if a traditional multiple-choice exam is given early in training, then it may be erroneous to conclude that the highest scorers on the exam are the ones most likely to continue to succeed in training, given that those who are the most intelligent will have an (early) edge in acquiring declarative knowledge and in displaying it on a more traditional measure.

**Measurement implications.** At one level, the implications for evaluating declarative knowledge vary little from the suggestions of Kirkpatrick (1976, 1987) or from how organizations typically measure learning. The acquisition of declarative knowledge can be assessed through multiple-choice, true–false, or free-recall exams. At another level, it should be recognized that there are two different approaches for assessing the amount of declarative knowledge held by trainees. Speed tests assess the number of items answered in a given time or the reaction time to any single item. Power tests assess the number of items answered correctly, given unlimited time. Ackerman and Humphreys (1990) noted that given similar content, speed and power tests actually measure different constructs—an argument supported by empirical studies (e.g., Lohman, 1979; Lord, 1956). Ackerman and Humphreys suggested that format choice should depend on the construct to be measured. Power tests measure the accuracy of stored information; such tests should be used when recall accuracy is valued or the consequences of errors are high. Speed tests measure the rate at which individuals can access knowledge; they are more appropriate when recall speed is the focal learning outcome.

In practice, many evaluations confound the two constructs, such as when the number of correct answers are counted on a timed test. Ackerman and Humphreys (1990) offered their own criticisms of such tests, and the interested reader is referred to their chapter for more details. For present purposes, the immediate implications are that when selecting tests to evaluate initial knowledge acquisition, thought should be given to the precise nature of the construct in question and that the choice of testing format should reflect that construct.

The second measurement implication is more direct. To maximize the value of traditional measures of declarative
knowledge, these tests should be given closer to the beginning of training than to the end. There are two reasons for this. First, as a feedback mechanism, such tests should be given early enough to identify knowledge gaps that may hinder the rate of subsequent (higher-order) learning. Second, from a psychometric perspective, variance among trainees in declarative knowledge should be greater earlier in training than near the end. Consequently, earlier evaluation scores would have the greatest use for predicting other learning outcomes.

**Knowledge Organization**

As skill learning advances beyond initial acquisition phases, several interrelated changes in processing occur as a function of continued practice (Anderson, 1982; Fitts & Posner, 1967). First, learners begin to focus less on declarative knowledge and more on procedural knowledge. Concurrent with an increase in procedural knowledge is the development of meaningful structures for organizing knowledge. In recent years, many researchers have stressed that of equal or greater importance than the amount or type of knowledge stored in memory is the organization of that knowledge (Johnson-Laird, 1983; Rouse & Morris, 1986).

The term *mental model* has been used to describe how people organize their knowledge. Mental models serve as mechanisms trainees use to describe functions and forms of tasks, explain and observe integration of tasks, and anticipate future task requirements (Rouse & Morris, 1986). Synonyms for mental models include knowledge structures, cognitive maps, or task schemata. Incumbents are believed to possess separate models for multiple functions on the job. For example, a military pilot may possess distinct mental models for preflight briefings, takeoffs, landings, tactical engagements, and aircrew coordination. Mental models provide a context for the interpretation of objects and events; they not only organize existing information, but influence the acquisition of new knowledge (Messick, 1984).

One important characteristic of mental models is the type or complexity of the stored elements. Studies of expert/novice differences suggest that whereas novices create different mental models for problem definition and solution strategies, experts form more complex knowledge structures that contain both problem definition nodes and solution nodes (Glaser & Chi, 1989). The advantage of this model is that having identified the problem, experts are able to quickly access a solution strategy because that strategy is closely linked (in memory) to the problem node. In contrast, solution times are slower and less fluid as the novices must engage in one search to identify the problem and another to solve it (Glaser, 1986).

A second important characteristic of mental models is the organization or interrelationships among model elements. Research on expert/novice differences reveals that experts’ knowledge bases are characterized by hierarchical storage and strong paths between critical elements (Glaser & Chi, 1989). *Hierarchical storage* refers to the way in which new information is integrated and the way in which existing knowledge is organized. Additionally, through experience, paths between diagnostic nodes and solution nodes are solidified, increasing solution speed.

**Measurement implications.** One direct implication of the previous discussion is that trainees’ understanding of course material may be best assessed by measuring their supportive cognitive structures. There are numerous strategies for directly measuring these structures (Flanagan, 1990). The modal technique requires judgments of similarity or closeness among previously defined sets of core elements. Elements are then mapped either by having the learners physically arrange the elements using a free-sort task within a problem space (e.g., Champagne, Klopf, Desena, & Squires, 1981) or by submitting the judgments to a clustering or scaling algorithm (e.g., Cooke & McDonald, 1987; Naveh-Benjamin, McKeachie, Lin, & Tucker, 1986; Shavelson, 1974). This latter strategy is referred to as *structural assessment* (Goldsmith & Johnson, 1990).

The resulting map can be “scored” by assessing its similarity to a prototype or to the instructor’s map (Goldsmith & Davenport, 1990) or by determining its level of complexity (Champagne et al., 1981). For example, the greater the number of levels or the greater the differentiation among elements within a level, the more complex the map. Other scorable elements of a mapped structure include counting the number of levels (as an indicator of hierarchical organization), determining the distance or number of links between key nodes (e.g., between diagnoses and solutions), or assessing overlap to a prototype after weighting links by their importance to job success or future development.

A good example of the use of mental models to evaluate training programs is found in the research of Goldsmith and Johnson (1990; Goldsmith, Johnson, & Acton, 1991). Goldsmith and Johnson assessed changes in the knowledge domain over a 16-week undergraduate course on psychological research methods. Working with the department faculty, they defined a list of 30 core concepts to be covered during the course (e.g., confound, design, error, interaction). The rating task presented students with all possible pairs of concepts and required them to judge the relatedness of the concepts using a 7-point scale. Forty subjects completed the ratings during the 1st, 8th, and 15th weeks of the semester. The rating task assumes that the less related students perceive two items, the further apart the concepts exist in their knowledge structures. The ratings were analyzed using a procedure called Pathfinder (Schvaneveldt, Durso, & Dearth, 1985). Pathfinder creates a link-weighted network, a configuration in which concepts are depicted as nodes and relationships are depicted as links between nodes. A similarity index was computed between each student’s representation and the instructor’s for each measurement occasion. The index is approximately equal to the number of common links between two networks divided by the total number of links in both (Goldsmith & Davenport, 1990).

Goldsmith and Johnson (1990) did not attempt to assess learning over the course of the semester, so there were no data comparing students’ structures from the beginning of the semester to the end. Of interest though are analyses of the validity of end-of-semester cognitive maps for predicting total semester points. The correlation between the similarity index and final course points was .74. Thus, the more a student’s mental representation of psychological research matched the instructor’s, the better the student did in the course.

More recently, the same technique has been applied in true
training settings. Kraiger and Salas (1992) administered both the Pathfinder technique and a more traditional (multiple-choice) exam at the end of two training programs—one on aircrew coordination training for Navy pilots and the other on SPSS programming to graduate students in psychology. Kraiger and Salas found that the similarity between trainees' cognitive maps and those of training experts (those who designed the training, delivered the training, or both) were correlated with the traditional knowledge-based measure for Navy pilots and with traditional measures and task-specific self-efficacy and later transfer of training to classroom performance for the graduate students.

Cognitive Strategies

A final category of cognitive measures, drawn from Gagne's (1984) learning objectives, centers on the development and application of cognitive strategies. Individual differences exist in the extent to which knowledge can be accessed or applied more rapidly or more fluidly (Anderson, 1982; Kanfer & Ackerman, 1989). Anderson's (1982) theory proposed that skill development is a continuous process, and as knowledge and procedures continue to be compiled, more elegant task strategies emerge. Kanfer and Ackerman's (1989) theory proposed that through continued practice, complex behaviors are internalized; the greater the internalization, the more cognitive resources are available for executive functions or strategy development. Given that these executive skills or strategies are evident primarily at the highest levels of skill acquisition, measures of such would serve as capstones for training evaluation.

The term metacognition has been used to refer to both the knowledge of one's own cognition and the regulation of such (Brown, 1975; Leonesio & Nelson, 1990). Metacognitive skills include planning, monitoring, and revising goal-appropriate behavior (Brown, Bransford, Ferrara, & Campione, 1983; Schoenfeld, 1985) or understanding the relationship between task demands and one's capabilities (Pressley, Snyder, Levin, Murray, & Ghatatala, 1987). They also include skills in regulating or evoking appropriate strategies (Bereiter & Scardamalia, 1985). Strategies refer to a broad range of mental activities that facilitate knowledge acquisition and application (Prawat, 1989).

Research in cognitive psychology has shown that the metacognitive skills of experts are superior to those of novices. In contrast to novices, experts are more likely to discontinue a problem-solving strategy that would ultimately prove to be unsuccessful (Larkin, 1983), are more accurate about judging the difficulty of new problems (Chi, Glaser, & Rees, 1982), and are better able to estimate the number of trials they will need to accomplish a task (Chi, 1987). Similarly, good readers are more aware than are poor readers whether they are comprehending text as they read it (Pressley et al., 1987). In sports psychology, evidence of tennis strategy and effective decision making has been found to be a precursor to motor skill development and enhanced performance (McPherson & French, 1991; McPherson & Thomas, 1989).

These findings are relevant in two ways. First, they indicate that metacognitive skills of awareness are correlates of cognitive skill development and, hence, suitable indices of learning. For example, one study by Peterson, Swing, Stark, and Waas (1984) found that measures of students' cognitive processes were better predictors of achievement test scores than were observers' ratings. Measures of awareness, self-evaluated learning, needed development, and so forth would all seem to hold a place in the training evaluation domain as evidence of learning.

Second, because metacognitive skills of self-regulation are important to successful task performance, they also warrant measurement. For example, safety trainers in a nuclear power plant would certainly be interested in knowing which trainees could detect their own errors and which could not. Here, the emphasis is on the regulatory functions (Nelson & Narens, 1990). Knowing trainee skill levels of self-regulation or self-control of cognitive processing may provide valuable information when making job assignments, reassessments to training, and so forth. Furthermore, there is some evidence that the lack of these skills can lead to "production deficiencies" in subsequent learning and problem solving (Hertel & Hardin, 1990). A production deficiency occurs when the incumbent has the necessary abilities and knowledge to perform the task but lacks the metacognitive skills that facilitate the access to and use of these resources (Sliie & Weaver, 1992).

Measurement implications There are numerous ways in which the measurement of metacognitive skills can support inferences of learning during training. One commonality of these approaches is that each attempts to test for higher levels of understanding than that shown by mere recall or recognition. Measures of understanding should be designed to assess trainees' awareness of steps undertaken or progress toward a goal. Gott and her associates (e.g., Glaser, Lesgold, & Gott, 1986; Means & Gott, 1988) developed a method called probed protocol analysis to assess trainees' understanding of their task behavior relative to a superordinate goal. Working with a subject matter expert, they first defined, in detail, the steps necessary to solve a problem (e.g., an electronic troubleshooting task). Trainees were asked to describe what they would do at each step. Metacognitive awareness was assessed by asking trainees prespecified probe questions at each step. Examples of such questions are: "Why would you run this test, and what would it mean if it fails?", "How does this test help you solve the problem?", and "Are there any alternatives to the test you just ran?" Responses to these probes indicate whether trainees were generating and testing hypotheses, operating under goals, or understood when or whether they were making progress toward a solution. If probes are written to reflect breakdowns in routines, or goal blockages, the protocol analysis technique can be extended to assess self-regulatory skills, that is, the trainees' skill at initiating corrective responses.

Gill, Gordon, Moore, and Barbera (1988) used a protocol analysis to investigate the effectiveness of an instructional videotape. Gill et al. compared the diagnostic sensitivity of a set of question probes to free recall. The researchers found that the probe strategy resulted in a better representation of students' knowledge structures than did the free-recall task and that the knowledge structures built using these probes were highly correlated (r = .73) with problem-solving strategies. Kraiger and Salas (1992) developed a paper-and-pencil version of Means and Gott's (1988) probed protocol analysis technique. Responses were collected from graduate students after a 3-hour SPSS, training course delivered during the third week of a mul-
tivariate statistics course. Responses were scored for thoroughness and understanding by three subject matter experts. Post-
training scores correlated .68 and .74 with two final exam SPSS, problems administered 12 weeks later.

Research on meta-cognition reinforces the value of measuring trainees' self-assessments of learning outcomes. A strategy suggested by Bloom and his colleagues is to use self-reports of awareness or understanding (Krathwohl et al., 1964). Self-re-
port measures should reflect trainees' awareness of knowledge gained. Questions about trainees' awareness of level of the pro-
ceduralization, degree of additional learning needed, or awareness of mistakes are appropriate. The extent to which trainees' answers coincided with empirical verification would indicate the extent to which they were developing expertise-like representations of the job. For example, studies in instructional psychology have found students' reports of attention and cognitive pro-
cessing to be valid predictors of academic achievement (Peter-
on, Swing, Braverman, & Buss, 1982; Peterson et al., 1984).

Similarly, Slife and Weaver (1992) reported that subjects' esti-
mates of the precision of their problem-solving efforts are rea-
sonable correlates of actual performance.

A good example of this type of approach can be found in a training study conducted by Fisk and Gallini (1989). Fisk and
Gallini attempted to compare a traditional training program (for learning base-5 arithmetic) to one based on conceptually
based information processing principles. Among their evalu-
ation measures was a test of Perceived Readiness for Examina-
tion Performance (PREP; Pressley et al., 1987). PREP requires examiners to predict how many items they expect to answer
correctly on a test. It assesses the ability to judge whether mate-
rial is adequately learned and understood for future applica-
tions. Fisk and Gallini found that subjects receiving the hypo-
thetically superior form of training scored higher on the PREP
and scored higher on more traditional measures of learning as
well.

The success of the self-assessment measures discussed above
suggests that learners may yield accurate portrayals of current
knowledge states, provided that the information has been pro-
perly elicited. This conclusion is supported by several reviews
documenting the validity of self-assessments for psychological
assessment (Shrauger & Osberg, 1981) and academic ability
(Mabe & West, 1982).

One clever example of the use of self-ratings as indices of
awareness can be found in a training study by Schendel and
Hagman (1982). These researchers sought to determine if indi-
viduals could estimate in advance how much refresher training
they would require to regain a certain level of proficiency. Sol-
diers were put through a training program on how to disassem-
ble and reassemble an M60 machine gun. Training was com-
pleted to a point of overlearning. After a period in which the
skill was not practiced or rehearsed, the soldiers were asked to
estimate the number of trials required to return to their original
levels of proficiency after the original training. The results
comparing the actual number of trials with the self-assessments
indicated that the soldiers were fairly accurate in their predic-
tions.

Skill-Based Learning Outcomes
The second category of learning outcomes concerns the de-
velopment of technical or motor skills. Characteristics of skill
development include a goal orientation and a linking of behav-
iors in a sequentially and hierarchically organized manner
(Weiss, 1990).

Traditionally, skill development has been evaluated by ob-
serving trainee performance in role plays (simulations at the
end of training) or in actual job behaviors. Behavioral observa-
tion may be an appropriate evaluation tool, provided the as-
essment strategy is developed in concert with a theoretical
conceptualization of skill development.

Theories of skill development generally posit three definable
stages: (a) initial skill acquisition, (b) skill compilation, and (c)
skill acquisition involves the transition from knowledge that is
declarative to knowledge that is procedural (Neves & Anderson,
1981). Procedural knowledge enables the reproduction of
trained behaviors. Compilation skills occur with continued
practice beyond initial successes at reproducing trained behav-
ior. Performance at this stage is characterized by faster, less
error-prone performance and by the integration of discrete
steps into a single act. With subsequent automaticity, individ-
uals not only perform tasks quickly but are able to maintain
parallel rather than successive processing of activities (e.g.,
Shift-
lin & Dumais, 1981). Individuals at this stage are more likely
to detect the appropriate situations for using a skill (Gagne,
1986) and to individualize skilled acts.

Although initial acquisition must precede higher order skill
development, it is during the latter steps that successful initial
learning is translated into adaptive skills. Therefore, in this sec-
tion, we focus on the learning outcomes evident at the compila-
tion and automaticity stage as they represent the higher levels
of skill development that are often the desired outcomes of skill-
oriented training programs.

Compilation
If advanced skills are defined by smooth, fast performance
(Gagne, 1984), then trainee behavior at the initial skill acquisi-
tion stage of development may be characterized as rudimentary
in nature. Trainees may reproduce trained behavior but only
through a heavy reliance on working memory and mental re-
hearsal of previously learned routines (Weiss, 1990). Accord-
ingly, performance is slow and the trainees' ability to attend to
task-irrelevant information or to react to novel task-relevant
stimuli is low.

In contrast, trainee performance at the compilation stage is
decidedly faster and more fluid. Errors are reduced, verbal re-
hearsal is eliminated, and behavior is more task-focused. Ac-

According to Anderson (1982), compilation is the result of two
interrelated processes: proceduralization and composition. Dur-
ing proceduralization, the trainee builds smaller, discrete
devices into a domain-specific production or routine. For
example, a computer programmer builds and learns to apply
different debugging strategies for different types of problems.
Composition begins simultaneously with proceduralization but
may continue after it. During composition, the trainee mentally
groups steps by linking successive (previously learned) proce-
dures into a more complex production. For example, whereas a
less advanced tennis player may have proceduralized separate
skills related to following through on shots and approaching the
net, after composition, the player executes both acts as a single, fluid behavior.

With compilation, individuals are in a better position to determine the situations in which trained skills are useful or not useful. Individuals also learn to apply newly learned behaviors to unique settings (generalization) and to modify existing skills depending on the situation (discrimination) (Anderson, 1982).

Measurement implications. In many domains, it would be useful to track trainees’ skill development to assess progress and to design or modify other training interventions. For example, knowing that trainees have reached the compilation stage may lead to the provision of greater opportunities for practice under constant mapping conditions to enable movement toward automaticity (Shiffrin & Schneider, 1977).

To adequately sample and track changes in compilation as a function of training requires the measurement of highly specific criteria (Smith, 1976) that reflect maximal performance (Cronbach, 1960). High specificity involves the observation and description of discrete behaviors of an individual. Maximal performance measures focus on what an individual can do rather than on typical performance.

One way to measure the development of compilation skills is through targeted behavioral observations (Komaki, Heinzman, & Lawson, 1980). This methodology requires timed observations randomly spaced and for a long enough period to draw valid inferences. Observation categories should reflect specific, anticipated changes related to learning objectives. For example, trainees can be asked to demonstrate their skills on learned tasks, while observers track criteria such as the frequency of desired or undesired behaviors, time to completion, steps necessary to complete a task, the sequencing of steps, or the number of task-related errors. Compilation can be inferred from evidence such as plateaus in the rate of increase of desired behaviors, a sustained decrease in the frequency of undesired behaviors or number of errors, a rapid increase in the completion of steps, the reordering or elimination of more trivial steps, or the appearance of synthesis across previously discrete steps.

Hands-on performance measurement is a second technique for assessing compilation. Both the Army (Campbell & Associates, 1990) and the Air Force (Hedge & Teachout, 1991) have developed hands-on tests. In these, a series of steps are identified as important for successfully completing a number of tasks. Individuals are then observed performing the tasks by trained observers who either record whether each step was taken (go/no-go) or provide an evaluation of the quality of step completion (pass/fail) (Borman & Hallam, 1991).

Two disadvantages of hands-on testing are the long time required to administer such tests and, for technical jobs, the necessity of assembling and disassembling costly equipment. An appealing alternative is a structured situational interview (Latham, Saari, Pursell, & Campion, 1980). This process permits detailed and targeted evaluation by asking trainees to state how they would perform or complete a task. For example, in the Air Force, airmen are shown a piece of equipment and asked to describe the steps necessary to complete a particular task. Results indicate that the situational interview is a reliable method for collecting information (Kraiger, 1991) and that performance on a situational interview correlates highly with performance on actual hands-on performance tests (Hedge & Teachout, 1991; Kraiger, 1991).

With either the hands-on or interview methodology, compilation can be assessed using many of the same criteria described under the behavioral observation methodology (measuring the frequency of undesired behavior, timing step completion, or watching for the elimination of steps). A more novel approach is to instruct the trainee to perform the hands-on task(s) twice—one under instructions to perform a task the way they would normally and a second time with instructions to perform the task as they were trained to do. Smoother, quicker, better performance with the use of fewer steps under “normal” conditions would permit inferences of compilation.

If the trainee is proficient at the task but unable to simulate the more mechanical, trained method, evidence exists that the trainee has passed through the compilation stage (Kraiger, 1988).

Compilation is also characterized by the capacity to modify learned behaviors to new task settings. This characteristic can be assessed by examining trainees’ ability to generalize skills beyond the situations trained and to discriminate when skills need to be adapted to fit a changing situation. For example, in behavioral modeling, trainees could be asked to attend to a situation by closely following the learning points from the training program across a variety of situations. In another scenario, trainees could be asked to adapt their trained behaviors to deal effectively with a unique situation that could be encountered on the job. Compilation skills would be evident when individuals are able to generalize skills to situations not specifically trained and when they are able to quickly and successfully adapt skills to unique situations with little effort or thought.

This approach was used in a study of salesperson effectiveness in customer relations (Leong, Busch, & John, 1989). In one scenario, salespeople had to recognize the similarities across various situations and respond by using trained skills. In a subsequent scenario, salespeople were presented with an atypical sales situation (customer characteristics did not match a trained profile). In this setting, the degree of compilation can be assessed from the extent to which the salespeople varied their selling approaches and tailored them to the individual needs of the customer. Those salespeople high in compilation skills should be able to incorporate more distinctive actions and be more adaptive in atypical situations.

Automaticity

Through continual practice, trainees may reach the automaticity stage. Although compilation and automaticity are most appropriately thought of as points on a continuum rather than discrete stages, there are certain characteristics of automatized behavior that are not evident during the compilation stage. The development of automaticity implies a shift in operational modes, from controlled to automatic processing (Schneider & Shiffrin, 1977; Shiffrin & Schneider, 1977). With automaticity, performance is fluid, accomplished, and individualized.

Because the learner can no longer verbalize intended behaviors or processes, automatization enables task accomplishment without conscious monitoring and enables concurrent performance on additional tasks (Shiffrin & Dumais, 1981). Further-
more, because attentional requirements decrease when behaviors are automatized (Ackerman, 1987), task behaviors are not affected by other demands for cognitive resources (e.g., distracting thoughts, situational pressures, secondary tasks, etc.). With automatization, individuals have greater cognitive resources available to cope with extraneous demands. Finally, during the automaticity stage, skill capabilities may be expected to undergo additional “tuning” (Rummelhart & Norman, 1978). Tuning involves changes such as improved accuracy, generalized applicability, specialized applicability, and determination of a prototype or typical case (Gagne, 1984). For example, most drivers have automatized important skills related to operating a car. This capacity allows them to converse with passengers while monitoring the road, changing speeds, or reacting to environmental changes.

Measurement implications. Collecting measures that assess automaticity also provides valuable information on the degree of trainee learning. Strategies effective for tracking compilation are not as applicable at this stage, given that automaticity is marked by significantly different cognitive and attentional processes.

Instead, automaticity would have to be assessed through measures specifically designed for that purpose. Cognitive psychologists typically have measured automaticity by examining artificial tasks (e.g., learning nonsense syllables) under rigorous experimental conditions that would not be found in applied contexts. This has led some writers to express doubt as to whether automaticity can be measured in organizational settings (Glaser et al., 1986). At least three strategies, however, hold some promise.

The first strategy directly parallels work by cognitive psychologists and requires trainees to perform the trained tasks while simultaneously performing a secondary task. Ideally, multiple measures on both tasks would be collected during training. Because automaticity minimizes cognitive effort or attention necessary to complete the primary task, as the trainee shifts from controlled to automatic processing, performance on the secondary task should increase. When performance on both tasks stabilizes (and reaches proficiency on the secondary task), automaticity may be inferred. For example, Tyler, Hertel, McCallum, and Ellis (1979) required subjects to unscramble anagrams such as “drootc” (which produces “doctor”). After each item was presented, a tone was delivered at one of four intervals through headphones worn by the subjects. Their secondary task was to monitor and respond as quickly as possible to the tone while solving the anagrams. One training analog for this task would be in the assessment of skill development in a plant operator. If trainees were instructed to gradually monitor more and more stimuli (e.g., meters, screens, etc.), stable performance on the primary task coupled with enhanced performance on the secondary task would suggest automaticity.

A second strategy is to use a single task but to have subjects solve both normal and interference problems. An interference problem resembles a normal one but with key information altered. For example, in math, an interference equation would be “4 + 5 = 20.” This equation is false as presented but could be true if the addition sign were changed to a multiplication sign (Zbrodoff & Logan, 1986). If automaticity has not occurred, interference problems (in this case, evaluating an equation as true or false) should not take longer, or be more difficult, than normal problems. As automaticity develops, the learner processes larger chunks of information and may not attend to discrete, though critical, pieces of information. Thus, performance decrements on interference tasks permit inferences of automaticity. Again, multiple measures during training will provide a baseline and lead to more confident judgments of the onset of automatization. Soloway, Adelson, and Ehrlich (1988) used such a task in a study of expert/novice differences in computer programming. Subjects were asked to provide a missing line in either a normal or “distractor” program. The distractor program ran successfully but contained elements that violated rules of programming discourse (e.g., a function that returned the minimum of a set of values was assigned to a variable named MAX). As expected, performance differences between proficient and nonproficient subjects were more pronounced on the distractor programs.

A third strategy is not to measure automaticity directly but to embed a concern for it in all measures collected during or after training (Glaser et al., 1986). Evaluation is focused on other dimensions of knowledge or behavior. However, variations in the context or methods of assessment may enable inferences of automatization. For example, suppose that pilot training is to be assessed using two behaviorally based simulated flights at the end of training. By constructing one scenario to be context-rich, with multiple interference tasks, and a second to be context-simple, automaticity is implied by the absence of performance differences between tasks. Another indication of automated processing would occur when trainees appear to stop monitoring their own behavior (or report less awareness of their actions). This observation may be confirmed by investigating whether trainees’ performance subsequently declines when they are explicitly asked to attend to their behavior. Other embedded measures suggested by Glaser et al. (1986) include determining whether goal structures exist independent of technical guidelines (e.g., Can a mechanic correct a malfunction without a repair manual?) or giving trainees a randomized list of steps to a task and asking them to select and order the important tasks. Here, the omission of initial steps (e.g., consult manual) could suggest automaticity.

Affectively Based Learning Outcomes

Gagne (1984) included attitudes as a learning outcome, reasoning that attitudes can determine behavior or performance, that there are a variety of mediums devoted to affecting attitudes (e.g., advertising), and that there is evidence that attitudes can be changed. Gagne defined an attitude as an internal state that influences the choice of personal action. Accepting this definition, we broadened this category of learning outcomes to include motivational and affective outcomes. We believe this treatment is consistent with Gagne’s classification because motivation is also an internal state that affects behavior.

An emphasis on behavioral or cognitive measurement at the expense of attitudinal and motivational measurement provides an incomplete profile of learning and the learning process (Gagne, 1984; Messick, 1984). Unfortunately, Kirkpatrick (1976, 1987) and others in the training field have ignored affectively based measures as indicators of learning. Instead, train-
ing researchers have collected what are termed reaction mea-
sures—indicators of how well trainees liked the training, per-
ceptions of how well organized the training program was, and
whether trainees found the training useful. These measures
provide feedback on the quality of training delivery but are not
direct measures of individual learning. Our intent in this sec-
tion is to propose a broader range of affectively or attitudinally
based outcomes that may be measured and used to infer learn-
ing during training.

Taking this broader perspective, we refer to affectively based
measures of training evaluation as a class of variables com-
prising issues such as attitudes, motivation, and goals that are
relevant to the objectives of the training program. In addition,
like Bloom and his associates (Bloom, 1956; Krathwohl et al.,
1964) and Gagne (1977), we conceptualize this class of variables
as indicators of learning, rather than simply precursors to learn-
ing. If a trainee's values have undergone some change as a func-
tion of training, then learning has occurred.

For purposes of classification, we have placed in this final
category all those learning outcomes that are neither cogni-
tively based nor skill based. These remaining outcomes are
generally of two types: those that target attitudes or preferences
as the focus of change and those in which motivational tenden-
cies are an indirect target of change.

**Attitudinal Outcomes**

The field is replete with examples of training programs that
establish affective change as a focus of training. The incorpora-
tion of affective outcomes into the training curriculum is most
clearly seen in the development of training programs (e.g., po-
lice recruits) that not only impart knowledge and skills but are
also powerful socialization agents (Feldman, 1989). Among the
affective outcomes in organizations that may be acquired by
training are creative individualism (defined as the acceptance
of pivotal norms and values and the rejection of all others;
Schein, 1968), organizational commitment (Louis, Posner, &
Powell, 1983), recognition of what is important to learn (Becker,
Geer, Hughes, & Strauss, 1961), group norms (Feld-
man, 1984), and tolerance for diversity (Geber, 1990).

Training objectives may also include desired outcomes such
as inner growth, self-awareness, and changing values. For exam-
ple, safety training programs often seek to influence the va-

ciences trainees attach to safe behaviors (e.g., Gregorich,
example involves one of the major training methods in the
field, behavioral modeling. What researchers and practitioners
often overlook is that this technique sets changing attitudes and
value systems as its ultimate goal. An examination of the theory
behind the behavioral modeling method reveals that skills
must first be developed through observation, practice, and rein-
forcement. Once the new behavioral patterns are found to be
effective in solving real problems in organizations, the individu-
als should come to recognize the value of those new and effect-
ive behaviors and then internalize attitudes and values con-
gruent with the new behavioral patterns (A. P. Goldstein &
Sorcher, 1974; Kraut, 1976).

**Measurement implications.** The achievement of outcomes
such as tolerance for diversity or concerns for safety can best be
measured by following principles for attitude measurement.

Measures of attitudes must take into account both the direction
of feelings toward the attitude object and the strength of the
reaction to the object. In the case of training evaluation, the
attitude object is defined by the specific learning objective(s).
The first focus of measurement should be whether the direction
of attraction toward the attitude object is consistent with the
learning objective(s). The direction of attitudes is assessed in
several ways. Most commonly, the attitude objects are listed
with a scale that allows the respondent to indicate a preference
or rejection of the object (e.g., an agree/disagree scale).

The second focus of measurement concerns attitude
strength, that is, how deeply an individual holds the attitude
(Chaiken & Stangor, 1987). There are a number of different
conceptualizations of attitude strength including attitude acces-
sibility (Fazio, 1988), attitude centrality (Kronsick, 1986), and
internalization or conviction (Abelson, 1988). Accessibility
refers to the number of cognitive associations between an object
and its evaluation. These associations are formed through the
individual actively processing information about an object. At-
titude centrality focuses on the degree of interconnectedness
between the individual and the attitude object. Internalization
or conviction focuses on the acceptance by the individual of the
attitudes, codes, principles, or sanctions that become a part of
an individual and that affect value judgments and determine
conduct (Krathwohl et al., 1964).

Abelson (1988) suggested a number of ways of using self-rep-
port measures to operationalize attitude strength. These in-
clude how strongly individuals hold to their views; how impor-
tant their views are to their self-perceptions; how concerned
they are about the issue in question; how often they think about
the issue; and how often they express views on the issue to
friends, co-workers, and family members. Other potential indi-
cators of attitude strength include steadfastness (how likely
they are to change their minds on an issue), affect (to what
extent do they feel angry or good when thinking about an issue),
certainty (how correct do they think their views are), and cen-
trality (how many other issues come up when discussing the
issue in question) (Abelson, 1988).

Once the learning objective is specified as the attitude object,
measures of attitude strength can be useful for inferring learn-
ing during training. Specifically, pre- and posttraining mea-
sures of attitude strength that indicate a change from confor-
mity and passivity to active participation and identification
with training goals would signal that learning has occurred.
Additionally, attitude strength may be expected to affect future
processing of information relevant to an existing attitude (Chai-
ken & Stangor, 1987). With training, those who become more
committed to an attitude or value are more likely to pay atten-
tion or process new information relevant to that value than
those who are simply passive or noncommittal to training goals.

Gregorich et al. (1990) provided one example of assessing
attitude change (although not specifically addressing attitude
strength) to evaluate a training program. The program con-
sisted of a seminar on cockpit resource management that was
delivered to captains, first officers, and flight engineers. The
researchers developed a Cockpit Management Attitudes Ques-
tionnaire that consists of 25 items chosen to measure attitudes
related to cockpit resource management. More than 500 flight
Captains, first officers, and flight engineers completed the questionnaire before and after training. The results indicated that there were significant changes in stated attitudes in the desired direction especially in relation to attitudes toward communication and coordination, command responsibility, and the recognition of stressor effects.

**Motivational Outcomes**

The second type of affective outcomes in training concern motivational tendencies. In some circumstances, changes in trainee motivation may be an intended outcome of training. A popular example of enhanced motivation as an intended outcome is a “free” real estate or money management seminar intended to increase attendees’ desire to attend (and pay for) additional training. More commonly, however, motivational change is a secondary training outcome. That is, although the primary objective may be to build skill sets, motivational changes are anticipated as well. In subsequent sections, we discuss three motivational outcomes that may be secondary objectives of training: motivational dispositions, self-efficacy, and goal setting. Each has been the focus of considerable research in other psychological disciplines, and each shows promise as a measure of secondary learning in training.

**Motivational disposition.** Research on cognitive development in children has led to the recognition of two distinct motivational dispositions within the classroom: mastery or performance orientations (Dweck, 1986; Dweck & Elliott, 1983). A *mastery orientation* is characterized by a concern for increasing one’s competence regarding the task at hand. A *performance orientation* is marked by an intention to do well and to gain a positive evaluation by others. A young baseball pitcher who tries to use his fastball to strike out all batters displays a performance orientation, whereas another who risks a bad outing to work on developing other pitches displays a mastery orientation.

People with a performance orientation perceive their capacities as fixed or immutable to change, whereas people adopting a mastery orientation perceive their skills and abilities as malleable and therefore establish learning goals to develop those qualities (Dweck & Leggett, 1988). A mastery orientation implies a tendency to make internal attributions for success and failure; to see learning as personally determined; and to view goals as flexible and individual, rather than normative and immediate (Prawat, 1989).

Although the performance and mastery orientation was originally conceived of as an individual difference variable (Dweck & Elliott, 1983), more recent research has characterized it as a general motivational tendency that is adaptable to situations or interventions. For example, researchers have shown that the best disposition depends on the material to be learned or the level of instruction (Ames & Archer, 1987; Biggs & Rihn, 1984) and that superior learners are able to adopt the orientation most appropriate to the context (Biggs & Rihn, 1984).

In a posttraining context, the best disposition may be a function of the nature of the task and the incumbents’ level of sophistication. Mentoring is one developmental activity that typically seeks to instill a mastery orientation in learners. For example, diagnosing illnesses from X-rays is a very difficult task to learn, requiring as many as 200,000 trials to achieve expertise (Lesgold et al., 1989). As in many other domains, the developmental pattern for radiological diagnosis begins with high levels of declarative knowledge, ingraining of repeatable patterns, and initial attempts at hypothesis formation and testing (Lesgold et al., 1989). Presumably, doctors faced with the task of training residents would attempt to instill a mastery orientation during the initial stages of skill development. Residents should be willing to make and learn from their mistakes. However, near the end of their residency, when the trainees assume greater responsibility for patient outcomes, a performance orientation (getting it right) should predominate. Periodic measurement of these orientations (e.g., Ames & Archer, 1987) would indicate the extent to which the training staff was affecting the proper motivation.

**Self-efficacy** Self-efficacy refers to one’s perceived performance capabilities for a specific activity (Bandura, 1977a). Self-efficacious perceptions are generally believed to be task-specific and are hypothesized to influence one’s choice of activities, effort expended, persistence, and task performance (Bandura, 1977a). Numerous studies have reported positive relationships between self-efficacy beliefs and task performance (e.g., Barling & Beattie, 1983; Stumpf, Brief, & Hartman, 1987; Weinberg, Gould, & Jackson, 1979). According to Bandura (1977a), psychological procedures such as training change behavior in part by creating and strengthening self-efficacy.

Accordingly, changes in trainees’ self-efficacy may be a useful indicator of learning or skill development during training. In some instances, enhanced self-efficacy will be a formal objective of the training. For example, sports psychologists believe performance expectancies and self-efficacy to be important determinants of athletic success (Feltz & Doyle, 1981; Feltz & Weiss, 1982). Consequently, numerous authors have advocated various coaching and training strategies to directly enhance athletes’ self-efficacy, including ensuring performance accomplishments (Weinberg et al., 1979) or providing various forms of modeling (Gould & Weiss, 1981; McAuley, 1985).

Alternatively, enhanced self-efficacy may be an unintended outcome of other, well-designed training programs. When difficult tasks are broken down into component parts (Blum & Naylor, 1968) and trainees are able to develop competency on simpler tasks before proceeding to more complex tasks (Briggs, 1968), the trainees are likely to develop stronger perceptions of self-efficacy concurrent with greater skill capacities. In a similar vein, Schunk (1985) postulated that as students work on a learning task, their perceived capabilities are influenced by such factors as their ability to cognitively process directions and instructions, early successes and failures, and other contextual factors such as rewards or social comparisons. Regardless of whether changes in self-efficacy are formal or unintentional training outcomes, it is clear that such changes occur and that they may be related to other training criteria of interest.

There are several reasons why perceptions of self-efficacy should be included as posttraining measures of learning. As stated before, enhanced self-efficacy may be a formal training objective. Additionally, such beliefs may moderate the relationship between knowledge acquisition and subsequent performance. It is a well-accepted fact that there is a difference between possessing relevant skills and applying them in diverse,
appropriate situations. Perceptions of self-efficacy may be one factor that determines whether or not trainees apply the skills they have acquired (cf. Bandura, 1983). Finally, posttraining self-efficacy beliefs may be useful predictors of long-term transfer or skill maintenance (Marx, 1982). For example, Kraiger and Salas (1992) reported that self-efficacy judgments collected at the end of training predicted scores on performance tests administered 3 months later better than a traditional test of learning did.

**Goal setting.** Considerable research has implicated the role of goals and goal setting in motivational processes (e.g., Locke & Latham, 1990; Naylor, Pritchard, & Ilgen, 1980). The mechanisms presumed to operate through goal setting are also those that characterize motivated behavior: direction, arousal, and persistence of effort (Locke & Latham, 1990). Research indicates that individuals who set specific, difficult goals and who are committed to those goals are more likely to exert effort and perform at a high level (Locke, Latham, & Erez, 1988; Mento, Steele, & Karren, 1987; Tubbs, 1986).

The importance of goal setting to training evaluation rests on three assertions. First, individuals differ in the extent to which they are active in self-management processes, including setting and working toward goals (Mann, 1986). Thus, it is reasonable to assume that, irrespective of an intervention or evaluation, there will be variability among trainees regarding the difficulty, complexity, or specificity of their goals for the application and development of recently acquired skills.

Second, individual differences in the type and structure of goals have been identified as an important difference between task experts and novices. This is important because training has been linked to a process of turning novices into experts (Howell & Cooke, 1989; Kraiger, 1988). Accordingly, it can be argued that difference in goal quality is a useful indicator of trainees’ development. Glaser (1986) noted that experts and novices differ in their goal clarity and specificity. In contrast to novices, experts are characterized by well-differentiated, hierarchical goal structures. Thus, trainees who have learned the most will exhibit better quality goals and can be expected to exert more effort in desirable directions, be able to work toward concurrent subgoals, and have contingency plans in the event of goal blockage.

There are other predictable differences between the goal structures of experts and novices (Glaser et al., 1986). For example, one common weakness of novice goal structures is that subgoals are nonindependent; for example, a new assistant professor may set conflicting first-year subgoals of winning a teaching award and establishing a broad-based research agenda. Additionally, novices also set goals based on known methods rather than on methods best suited for the problem at hand. Thus, the new faculty member may set out to win the teaching award by using lecturing as the predominant instructional method, regardless of whether it is the most appropriate format for the specific learning objectives of each class.

Finally, it is asserted that individual differences in the presence and quality of goals may hold additional implications for the extent to which knowledge and skills acquired in training are applied to the job. Goal setting is viewed as an important factor influencing transfer of learned behaviors to the job (Baldwin & Ford, 1988; Wexley & Baldwin, 1986). To illustrate, a study by Farrell and Dweck (1985; cited in Dweck & Leggett, 1988) differentiated students who held learning goals (of acquiring new skills or extending their mastery) from students with performance goals (of establishing the adequacy of their ability or avoiding giving evidence of inadequacy or both). Students who held learning goals were more likely to actively attempt transfer tasks and scored significantly higher on such tasks. Not surprisingly, a strong goal-setting component is also included in relapse prevention strategies (Marx, 1982). Numerous studies have supported the hypothesis that awareness of goal setting increases the likelihood that knowledge and skills acquired in training are applied on the job (Frayne & Latham, 1987; Gist, Bavetta, & Stevens, 1990; Latham & Frayne, 1989; Wexley & Baldwin, 1986; Wexley & Nemeroth, 1975).

**Measurement implications.** A direct implication of the preceding discussion is that changes in trainees’ motivational states (motivational disposition, self-efficacy, and goals) may be inferred as evidence of their development during training. Various techniques exist for measuring these motivational states. The standard procedure for assessing the magnitude and strength of self-efficacy perceptions is straightforward (Bandura, 1977b). Trainees are given a set of specific performance outcomes of increasing difficulty and asked to indicate whether they are capable of accomplishing the objective and, if so, their confidence in doing so (see Gist, 1989, for an example).

Several direct measures of mastery and performance orientations exist as well (Ames & Archer, 1987; Nimmer, 1991). Nimmer’s measure assesses both goal and affective dimensions of the mastery orientation, has been validated on a college-age sample, and includes items such as “When I approach a new task, I am usually confident in my ability to master the task” and “My enjoyment with working on tasks tends to be overshadowed by my concern over others’ evaluations.”

Trainees’ goals may be measured by direct assessment of their goal commitment (e.g., Earley & Kanfer, 1985; Latham & Steele, 1983). Goal commitment is perceived as a reasonable surrogate of goal acceptance (Locke et al., 1988). A more sensitive measure of training success may be to assess trainees’ goal structures. For a variety of tasks, but particularly those involving problem solving or decision making, trainees can be asked to demonstrate their goal structures using methods such as free recall, focused interviews, and think aloud verbal protocols. The similarity between trainees’ structures and those of job experts can be assessed and used to infer training success. Dimensions on which similarity can be assessed include goal complexity, goal specificity, goal difficulty, and contingency planning. Alternatively, goal structures can be measured and compared using methods similar to those used to assess cognitive structures (Naveh-Benjamin et al., 1986; Schaneveldt et al., 1985; Shavelson, 1974).

Generally, main effects for training on motivational outcomes may be expected whenever the training was intended to affect those states. That is, in many instances, there should be an a priori learning objective targeting the motivational outcome to guide the measurement process. In such instances, pre- and postevaluations are critical given that variables such as mastery and performance orientation and self-efficacy are often conceptualized as individual difference variables. When trainees’ motivational states change as intended, learning dur-
ing training may be inferred, and subsequent posttraining evaluation may not be necessary to draw conclusions regarding the success of the intervention.

For example, Gist (1989) used changes in self-efficacy perceptions to compare two methods for training managers in idea generation. The two training formats were compared on a measure of self-efficacy and on the quantity and quality of trainees' performance on two idea-generation tasks. Consistent with hypotheses, the cognitive-modeling strategy (Meichenbaum, 1975) resulted in both greater self-efficacy and task performance. Moreover, within training groups, self-efficacy was significantly correlated to task performance.

In other contexts, changes in motivational states may be an unintended consequence of training programs. Schunk (1985) argued that reinforcement associated with the routine, administrative aspects of classrooms may trigger increases in students' self-efficacy that generalize to trained tasks. In addition, some trainees may spontaneously set difficult, challenging goals whereas others may not. Measures of unintentional motivational change may be useful not only as immediate measures of training success but as potential moderators of transfer of training or skill maintenance. Thus, one trainee with a low level of knowledge acquisition during training but a strong mastery orientation or feelings of self-efficacy may show strong distal transfer, whereas another trainee who scores high on a measure of declarative knowledge but lacks well-articulated transfer and application goals may not demonstrate long-term skill maintenance.

Discussion

Both the theory and practice of training evaluation have been hampered by the absence of a thorough, conceptually based model of training evaluation. We have suggested that this is in large part due to a reliance on incomplete theories of what learning is and how it should be evaluated. Accordingly, we discussed learning as a function of changes in cognitive, skill-based, and affective states. For each state, we reviewed relevant research from other psychological domains, proposed potential changes as a function of training, and recommended alternative methods for evaluating learning in training.

A Classification Scheme

As a result of these efforts, we are in a position to derive a preliminary, conceptually based classification scheme of learning outcomes for training evaluation. This scheme, presented in Table 1, is organized by the cognitive, skill-based, and affective learning outcomes. For each learning outcome, the table describes relevant learning constructs, appropriate foci for measurement, and potential evaluation measures. The classification scheme in Table 1 summarizes the constructs and measures we have proposed throughout this article. The scheme advances the training evaluation by providing a broadened definition of learning, a construct-oriented approach to learning outcomes, and the rudiments of a nomological network necessary to build models of training evaluation and training effectiveness. These contributions are clarified below.

Definition of learning Traditionally, industrial/organizational (I/O) psychologists have used definitions of learning such as "a relatively permanent change in behavior that occurs as a result of practice" (Wexley & Latham, 1991, p. 73) while ignoring more multifaceted definitions (Gagne, 1984; Messick, 1984). In contrast, and as indicated in Table 1, training and training evaluation can target a number of learning constructs, including declarative knowledge relevant to valued skill; development of complex and useful mental models for storing, organizing, and applying knowledge; development of strategies and executive functions for monitoring and regulating skilled performance; development of compilation skills such as proceduralization and composition; development of fluidity or automaticity in retaining and accessing knowledge; development and internalization of appropriate attitudes toward the focus of instruction; and changes in motivational tendencies.

It should be noted that these learning outcomes are not discrete but are often interrelated; that is, changes in one learning outcome may imply changes in another. For example, as trainees build accurate mental models, they must understand the task, equipment, co-workers, and the interrelationships among these elements (Cannon-Bowers, Salas, & Converse, in press; Rouse, Cannon-Bowers, & Salas, in press). In turn, these models lead to expectations about the elements and their interrelationships (Rouse et al., in press). These expectations fall closer to the affective domain than to the cognitive realm because they allow the learner to "sense" the correctness of current situations or "anticipate" nontrained phenomena.

It is evident from Table 1 that the adoption of a broader, multidimensional definition of learning leads to the consideration of a broader range of conceptually based evaluation measures. We have attempted to lay the foundation for a more comprehensive training evaluation model by adapting measures of various psychological constructs from other disciplines. The evaluation measures that we have suggested include power or speed tests of recognition or recall of declarative knowledge; knowledge elicitation techniques for measuring mental models such as verbal protocol analyses or cognitive mapping methods such as PathFinder; probed protocol analysis and self-reports of awareness and readiness to perform as measures of metacognitive strategies; alternative behavioral sampling methodologies for measuring compilation such as hands-on testing, structured interview testing, and behavioral observations; embedded tests for automaticity such as the use of distractor tasks; and perceptually based measures such as the assessment of attitude strength or goal complexity and pre- and postcomparisons of self-efficacy or motivational dispositions. In some instances, our recommended measures have yet to be used in an actual training environment. However, the success of these procedures for other purposes in other domains yields optimism that they would be similarly successful as mechanisms for training evaluation.

Construct-oriented approach Throughout this article, we have advocated a construct-oriented approach to the conceptualization of learning states and to the development of training evaluation measures. We have used the term learning construct to connote both a final state (e.g., proficiency, declarative knowledge) and a process for achieving that state (e.g., knowledge organization).

A construct-oriented approach to training evaluation has two
Table 1
*A Classification Scheme for Learning Outcomes for Training Evaluation*

<table>
<thead>
<tr>
<th>Category</th>
<th>Learning construct(s)</th>
<th>Focus of measurement</th>
<th>Potential training evaluation methods</th>
</tr>
</thead>
<tbody>
<tr>
<td>Cognitive outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Verbal knowledge</td>
<td>Declarative knowledge</td>
<td>Amount of knowledge</td>
<td>Recognition and recall tests</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Accuracy of recall</td>
<td>Power tests</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Speed, accessibility of knowledge</td>
<td>Speed tests</td>
</tr>
<tr>
<td>Knowledge organization</td>
<td>Mental models</td>
<td>Similarity to ideal</td>
<td>Free sorts</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Interrelationships of elements</td>
<td>Structural assessment (e.g., Pathfinder)</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Hierarchical ordering</td>
<td></td>
</tr>
<tr>
<td>Cognitive strategies</td>
<td>Self-insight</td>
<td>Self-awareness</td>
<td>Probed protocol analysis</td>
</tr>
<tr>
<td></td>
<td>Metacognitive skills</td>
<td>Self-regulation</td>
<td>Self-report</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Readiness for testing</td>
</tr>
<tr>
<td>Skill-based outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Compilation</td>
<td>Composition</td>
<td>Speed of performance</td>
<td>Targeted behavioral observation</td>
</tr>
<tr>
<td></td>
<td>Proceduralization</td>
<td>Fluidity of performance</td>
<td>Hands-on testing</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Error rates</td>
<td>Structured situational interviews</td>
</tr>
<tr>
<td></td>
<td></td>
<td>Chunking</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Generalization</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Discrimination</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Strengthening</td>
<td></td>
</tr>
<tr>
<td>Automaticity</td>
<td>Automatic processing</td>
<td>Attentional requirements</td>
<td>Secondary task performance</td>
</tr>
<tr>
<td></td>
<td>Tuning</td>
<td>Available cognitive resources</td>
<td>Interference problems</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Embedded measurement</td>
</tr>
<tr>
<td>Affective outcomes</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Attitudinal</td>
<td>Targeted object</td>
<td>Attitude direction</td>
<td>Self-report measures</td>
</tr>
<tr>
<td></td>
<td>(e.g., safety</td>
<td>Attitude strength</td>
<td></td>
</tr>
<tr>
<td></td>
<td>awareness)</td>
<td>Accessibility</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Centrality</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Conviction</td>
<td></td>
</tr>
<tr>
<td>Motivation</td>
<td>Motivational</td>
<td>Mastery versus performance</td>
<td>Self-report measures</td>
</tr>
<tr>
<td></td>
<td>disposition</td>
<td>orientations</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Appropriateness of orientation</td>
<td></td>
</tr>
<tr>
<td></td>
<td>Self-efficacy</td>
<td>Perceived performance</td>
<td>Self-report measures</td>
</tr>
<tr>
<td></td>
<td>Goal setting</td>
<td>capability</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Level of goals</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Complexity of goal</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>structures</td>
<td></td>
</tr>
<tr>
<td></td>
<td></td>
<td>Goal commitment</td>
<td></td>
</tr>
</tbody>
</table>

clear benefits. First, it forces researchers to be explicit in identifying not only the instructional objectives (e.g., specific knowledge, skills, and abilities) but also the most appropriate mechanisms for facilitating trainee development toward those objectives. Although in this article we have focused on the evaluation of training, implicit in this discussion is the need to be more explicit about the intended changes of training. Depending on the anticipated learning outcome, evidence of the success of the training program may be derived from mean differences (between pre- and posttests or trained and untrained groups), measures of convergence to a prototype, or measures of variability across trainees. For example, the compilation and automaticity constructs make different predictions about the degree of consistency across trainees during skill development. Thus, over repeated measures of skill-based testing, evidence of reduced variability over response patterns would support inferences of compilation, whereas subsequent evidence of increased variability would support inferences of automaticity (as trainees individualize performance).

A second benefit of a construct-oriented approach concerns the potential for validating the actual training measures used. Researchers typically have shown less interest in the construct validation of criterion measures than of predictors (Binning & Barrett, 1989; Schmitt, 1989). This orientation toward predictor constructs has been attributed both to a lack of clarity regarding the requirements of construct validation of performance measures (Austin, Villanova, Kane, & Bernardin, 1991) and to a lack of agreement regarding the nature or scope of
performance constructs (cf. Borman, 1991; Campbell, Campbell, & Associates, 1988; Fleishman & Quaintance, 1984). Although the identification and estimation of both relevant and irrelevant sources of variance remains a vexing problem in performance appraisal (Waldman & Spangler, 1989), these same issues are potentially solvable in the context of training evaluation. Given that the performance domain is more narrowly defined through learning outcomes, measured variables can be more readily linked to latent variables (intended learning outcomes), and extraneous sources of variance can be controlled or measured through the experimental or quasi-experimental designs commonly used in training evaluation.

**Development of a nomological network.** Other researchers have likened construct validation to theory development (cf. Clark, 1983). This comparison is relevant to the present case because we aspire to a theory or model of training evaluation. As a first step, we have developed a conceptual scheme to organize relevant learning outcomes, have identified appropriate learning constructs within this conceptual framework, and have specified multiple measurement techniques that can be applied to specific learning outcomes.

The next step is to identify a nomological network of multiple concepts, measures, and their interrelationships (Cronbach & Meehl, 1955). A nomological network requires the development and testing of hypotheses about the interrelationships among the learning constructs identified in Table 1, the relations of the learning constructs with other training outcomes such as behavioral changes on the job, and the predictive impact of factors such as trainee characteristics and training design on learning outcomes. Each step in building a nomological network for learning outcomes is described below.

An implicit assumption of the conceptual scheme presented in Table 1 is that there is a complex set of relationships among the measures of the various learning constructs. Several principles can guide the development of hypotheses for testing the interrelationships among these measures. First, stronger relationships should be expected among measures of the same learning construct than among measures of different constructs. That is, there should be evidence of convergent validity among multiple measures of the same learning construct, but there should be evidence of discriminant validity between measures of different learning constructs. More generally, operational measures of sequential learning constructs should be more highly correlated than measures of distal constructs. For example, measures of verbal knowledge should be more highly correlated with measures of either knowledge organization or compilation than with measures of metacognitive skills or automaticity. Furthermore, because metacognitive skills and automaticity are both indicative of higher order skill development, measures of these constructs should be highly intercorrelated.

A second step is to examine the relationship between changes in learning outcomes and other important training outcomes. As noted by Alliger and Janak (1989), Kirkpatrick's (1976, 1987) four steps of reaction, learning, behavior, and results have been codified in the training literature as a hierarchical approach to evaluation. For example, it is now accepted that learning is a necessary but not sufficient condition for behavioral changes on the job (e.g., see Clement, 1982; Noe & Schmitt, 1986). Alliger and Janak questioned this hierarchical approach and presented an alternative causal, plausible ordering. In their thinking, reactions to training are unrelated to learning, and learning can have a direct influence on results criteria as opposed to the indirect results it has on behavioral change.

The problem is not which causal ordering is preferable. Rather, from a construct validation perspective, the problem is that linkages have been identified without articulating reasons why or how such measures are logically related. Furthermore, a more substantive issue has rarely been addressed—under what conditions should different learning constructs be related to different behavioral change measures such as skill generalizability and maintenance. For example, it is likely that trainees who develop compilation skills will generalize training to more appropriate job situations than will those at the initial skill acquisition stage. Similarly, trainees who develop skills at the automaticity stage should maintain those skills over a longer period on the job than will those who leave training at the compilation stage. Relating learning criteria to intended learning outcomes enables the identification of such conditions.

A third step in building a nomological network is to specify and test relationships between the learning measures and predictor variables such as trainee characteristics and training design characteristics. Recent research on training characteristics and learning outcomes provides directions for such research (e.g., Tannenbaum et al., 1991). For example, Ackerman (see Ackerman & Humphreys, 1990) has shown that general intelligence is a good predictor of individual differences at the initial skill acquisition stage but is less of a predictor for trainees at the automaticity stage. Kanfer and Ackerman (1989) conducted three studies that indicated the complex interactions of type of training (e.g., declarative or procedural). These studies focused on ability differences, goal setting, and self-regulatory activities. For example, one study showed that goal-setting manipulations during initial training led to more self-regulatory activities and higher levels of performance.

Generally, learning measures would be expected to show the greatest effects when there is congruence between the learning objectives, instructional designs, and the method of assessment. For example, Glaser (1986) discussed several training techniques for improving learners' mental models. It should be expected that operationalizations of learning constructs relevant to mental models (e.g., similarity to an ideal) should show greater pre- to posttest changes when improving or building mental models are instructional objectives and when these mental modeling training techniques are used as opposed to the more traditional lecture and discussion methods of instruction. Similarly, measures of affective or motivational change should show the strongest effects in programs that make these outcomes an explicit training objective.

**Conclusion**

In this article, we have integrated theory and research from a number of diverse disciplines and have provided a multidimensional perspective to learning outcomes. We have advanced the theory of training evaluation by providing a conceptually based scheme of learning constructs, measurement foci, and measurement techniques. The value of our construct-oriented approach
is that it provides a systematic framework for conducting training evaluation research. The ultimate criterion for such work is whether it spurs additional research in training that advances our understanding of training evaluation and training effectiveness.

References


presented at the sixth annual meeting of the Society for Industrial and Organizational Psychology, St. Louis, MO.


Pressley, M., Snyder, B. S., Levin, J. R., Murray, H. G., & Ghatala, E. S.
K. KRAIGER, J. FORD, AND E. SALAS


Received September 20, 1991
Revision received June 29, 1992
Accepted June 30, 1992