1997.
In. M.A. Quiñones & A. Ehrenstein (Eds.), Training for 21st Century Technology: Applications of Psychological Research (pp. 63-88). Washington, DC: American Psychological Association.

3 Evaluating Training *During* Training: Obstacles and Opportunities

Dina Ghodsian, Robert A. Bjork, and Aaron S. Benjamin

The conditions of training can be manipulated in a great variety of ways, producing multiple possible configurations of a given training program. Trainees' posttraining performance, in turn, is heavily influenced by such manipulations: Some configurations of the conditions of training are far better than others. It is of obvious importance to choose those configurations that optimize the conditions of training, but doing so depends on accurate assessment, which itself can pose formidable difficulties.

One problem is that trainees' performance *during* training is an unreliable indicator of posttraining performance. Manipulations that enhance performance during training can yield poor long-term posttraining performance, and other manipulations that seem to create difficulties and slow the rate of learning can be optimal in terms of long-term performance (see Christina & Bjork, 1991; Schmidt & Bjork, 1992). Another problem is that trainees' own subjective evaluations of their knowledge and capabilities can be misguided, leading them to prefer nonoptimal training regimens (see Bjork, 1994b; Jacoby, Bjork, & Kelley, 1994).

Given those problems, it may seem obvious that one should assess training programs in terms of trainees' posttraining performance in the real-world settings that are the target of training. Optimizing on-the-job performance is, after all, the goal of typical training programs. For a variety of institutional reasons, however, assessing on-the-job performance in a meaningful way is often impractical, if not impossible.

It is the mission of this chapter to consider a question of theoretical

interest and practical importance: Are there some measures of performance obtainable during training that might serve as viable indicators of the degree to which the long-term goals of training are being met? In part to motivate the consideration of such measures, should they exist, we first summarize the difficulties and disadvantages of trying to use on-the-job performance as a measure of training. We then summarize the reasons why typical measures of in-training performance, both objective and subjective, are unreliable indicators of learning. We conclude by proposing some innovations in training programs that might provide more reliable measures of learning.

Impediments to Posttraining Assessment

Posttraining measures of on-the-job performance are potentially the most valid measures of a training program's effectiveness, but in practice, obtaining such measures is often difficult. There are both practical difficulties and institutional impediments that can preclude the collection or timely acquisition of posttraining data. We summarize a few of those difficulties and impediments in this section. (For a more complete discussion, see Bjork, 1994a.)

Practical Difficulties. In some cases, training is administered by an independent entity or by a training division so isolated from the rest of the organization that trainers simply do not have access to trainees once training is completed. In other cases, the lag between completion of the training program and receipt of posttraining data by training personnel is too long for the information to be of use in modifying the training regimen. Further, organizational units responsible for posttraining operations may lack the staff and resources necessary for extensive on-the-job assessment, or they may be deterred by the sensitive nature of such assessment. Testing in the work environment can cause significant apprehension by those being evaluated and may hinder normal operations.

Institutional Impediments. In addition to such relatively passive deterrents of adequate posttraining assessment, there are a number of more active institutional impediments. Often, there is resistance to the collection of field data, driven by a fear of liability. If an analysis of performance on the job reveals inadequate training, the training institution could face legal and financial penalties. Hence, management settles on a policy of ignorance. Another source of resistance to the mea-

surement of posttraining proficiency relates to the cost of retraining. If it is determined that an individual is performing at a substandard level, there is a pressure to retrain (or release) that individual. However, each time a person is sent back for retraining, there are costs to the organization. Again, in such an instance, the solution settled on is simply to remain unaware of the deficiency, under the assumption that what you don't know will probably hurt you less than what you might find out.

The numerous financial, organizational, and political deterrents just enumerated can eliminate posttraining assessment as a practical option in many settings. The other option is to assess training *during* training. As mentioned earlier, however, current methods of in-training assessment often provide unreliable indicators of long-term performance. Nevertheless, we believe that it is worth looking again at the potential for in-training assessment in the light of certain salient characteristics of human learning and memory.

Obstacles to In-Training Assessment

Individuals attempting to assess a training program's effectiveness can draw on two categories of measures—objective measures and subjective measures. *Objective measures* include tests of performance during training and tests administered at the end of training. *Subjective measures* include trainees' evaluations of training, often assessed on "smile sheets" or "happy sheets" administered at the end of training, and a trainer's own sense of satisfaction with the program (see Goldstein, 1993, for a discussion of evaluation techniques). Both objective and subjective measures of training effectiveness can be flawed indicators. They can be tainted or misinterpreted in ways that may lead to dramatic errors in assessment.

Interpreting Objective Performance

Of primary importance in the evaluation of the efficacy of a particular training regimen is assessing the *objective* state of a trainee's learning. Because performance on seemingly "objective" tests during learning can be biased by factors that will not be influential at a later time or in different circumstances, it is important to delineate those confounds known to obscure the relationship between actual long-term learning and performance on tests administered during learning.

Learning Versus Performance

Probably the most fundamental oversight on the part of a trainer is a failure to recognize the distinction between observed performance and actual learning. *Learning* is meant to refer to relatively permanent changes underlying behavior, brought about by manipulations of the conditions of training. The overt characteristics of behavior, which at a given time may or may not reflect such permanent changes, are referred to as *performance.* As Schmidt (1988) pointed out, performance during training can be propped up or impaired by short-term factors that are unique to the conditions of training and that may mask the actual level of learning that has been achieved.

One simple example is the effect of fatigue. Several studies have shown that although fatigue may depress performance during the acquisition of a skill, its effects on learning—as measured on delayed tests of retention—are frequently negligible (e.g., Alderman, 1965; Carron, 1969; Cotten, Thomas, Spieth, & Biasiotto, 1972; Schmidt, 1969; for a review, see Chamberlin & Lee, 1993). Schmidt (1969), for example, investigated the effect of fatigue on learning a ladder-climbing skill. All research participants performed a series of trials that involved climbing and balancing a free-standing ladder with an unfamiliar spacing between rungs. They were instructed to climb the ladder until it toppled over, as many times as possible in ten 30-second trials. Between trials, participants rode a stationary bicycle. The intensity of the exercise between trials was varied to induce different levels of fatigue. Whereas performance on the ladder-climbing task suffered with increased fatigue during the acquisition of the skill, there was no significant effect of fatigue on later ladder-climbing performance as measured by a retention test administered 2 days after the end of training. This experiment and others suggest that in the absence of knowledge about retention performance, an instructor observing the training performance of fatigued learners could easily and erroneously deem the training ineffective. It is thus crucial for an instructor to recognize that a given manipulation of the conditions of training can have vastly different short-term and long-term consequences.

Short-Term Versus Long-Term Consequences of Training

Just as fatigue acts to depress performance levels temporarily, other factors can artificially enhance training performance. Massing practice in a relatively short period of time often results in rapid improvement and high levels of performance during training (see, e.g., Bahrick, 1979;

Estes, 1955). Trainers are encouraged by such impressive results and can thereby be shaped into providing massed practice to keep performance levels high. Also, dividing a to-be-learned task into subtasks, which can seem a good idea to trainers, can result in massed practice on one subtask before the next subtask is introduced. For these reasons, there is a motivation for those responsible for training to implement massed training sessions. This motivation rests on the assumption—often false—that high levels of performance during training reflect high levels of learning.

There is evidence that during massed practice, learners can use the multiple, immediate repetitions to bypass some of the processes normally involved in producing a behavior: They simply repeat their performance from previous attempts (see, e.g., Jacoby, 1978). As a consequence, performance appears good during training, but little learning is actually achieved. In fact, after massed practice, performance tends to fall dramatically over periods of disuse. The opposite effect results from spaced practice. Trainees following a spaced practice schedule usually look worse than their counterparts engaging in massed practice during training but show significantly higher levels of retention at a delay (see, e.g., Baddeley & Longman, 1978; Bahrick, 1979; and Melton, 1970).

If a trainer can recognize that high levels of performance do not necessarily imply a high degree of learning, the goals of training can be shifted to a deeper level, at which future capabilities are given precedence over present functioning. Indeed, research on learning conducted in multiple domains suggests that instructors should introduce "desirable difficulties" for the learner during training in order to enhance long-term retention and transfer of skills to novel situations (Bjork, 1994b). Some examples of practice manipulations that depress performance during training but lead to a high level of long-term retention are discussed in the next paragraphs. (Also see Schmidt & Bjork, 1992.)

Reduced Feedback Frequency. One particularly counterintuitive learning effect is the effect of reducing the frequency of feedback provided to the learner. Here, the finding is that individuals who receive feedback about their performance after every attempt during training tend to perform more poorly on subsequent tests of retention than do individuals receiving less frequent feedback. Of course, there are limits to this generalization, as it is clear that some feedback is almost always better than no feedback.

The clearest evidence for an effect of reduced feedback frequency comes from motor-learning experiments, where feedback is often termed KR (knowledge of results). In an extensive review of the early literature on KR, Salmoni, Schmidt, and Walter (1984) noted the opposite effects of reduced feedback on performance during the acquisition of a skill and on retention performance. Winstein and Schmidt (1990) subsequently reported a set of experiments that nicely illustrated the effect. They used a horizontal lever moving task, in which individuals were to reproduce as accurately as possible a goal trajectory and movement time. They showed that reduced feedback at acquisition resulted in better performance on a no-KR retention test than did 100% feedback. Perhaps more interesting to note, however, was that the same effect was obtained when the retention test was conducted under conditions of 100% feedback (Experiment 3). That is, the benefits of reduced feedback frequency seemed to operate independently of the superficial similarity between acquisition and test conditions.

In a somewhat more applied study, Schooler and Anderson (1990) studied the effects of reducing feedback in the context of learning the computer language LISP. Training was to a large extent self-paced, and in addition to the enforced feedback differences between groups (high versus low frequency), there was some variation in the frequency of feedback administered within groups, depending on the specific nature of the errors made by individuals. Under these more relaxed conditions, retention was still facilitated by a decrease in the number of feedback presentations.

Variable Practice. Another effect that is robust across multiple domains is the advantage of variable practice. Experience with several versions of a task or materials during practice (variable practice), as opposed to only one version (constant practice), is often advantageous for learning as measured on a subsequent retention or transfer test. One characteristic of variable practice less marked in the other practice manipulations discussed here is that it results in a greater capability to generalize the knowledge or skills acquired during training to novel situations (see Lee, Magill, & Weeks, 1985; Van Rossum, 1990, for reviews).

The effects of variable practice are illustrated in an experiment by Catalano and Kleiner (1984). Participants in their experiment were seated perpendicular to a column of lights that, when illuminated in sequence, simulated movement toward the individual. The participants' task was to push a button coincident with the arrival of the "moving"

lights. Participants in the variable–practice condition performed ten trials at each of four different light speeds—5, 7, 9, and 11 mph—whereas participants in the constant-practice condition performed all 40 training trials at one of those four speeds. The transfer test consisted of five trials at either 1, 3, 13, or 15 mph (all outside the range of speeds experienced during training). Whereas during training participants in the constant group showed lower error scores than participants in the variable group, the reverse was true on the transfer test. On that test, participants in the variable group produced a significantly lower mean error score than participants in the constant group, supporting the claim that variable practice enhances a learner's capability to generalize from specific training tasks to novel conditions.

There is also some evidence that it may be beneficial to induce more variability during training than one expects to encounter in the setting that is the target of that training (see Bjork, 1994b). Consider a study conducted by Kerr and Booth (1978), who used children as participants. They asked 8-year-old children to toss miniature beanbags to a target. Children in the variable group practiced throwing from 2 ft and 4 ft, and children in the constant group always practiced 3 ft from the target. After completing the same number of training trials, children in both groups were given a test in which they were required to toss the beanbag at a target 3 ft away. Children in the variable group produced smaller error scores than did children in the constant group. This result is striking, because the criterion distance was exactly the distance (3 ft) at which the constant-practice group, by conventional wisdom, should have excelled. In this experiment and many others, however, conventional wisdom is a poor guide; the advantages of variable practice, such as learning to modulate the distance of a throw, apparently more than compensated for any specific practice advantage that accrued to the constant group.

Random Practice. The random practice effect is an effect of task-ordering within training sessions. In a typical experiment, one group of participants practices a set of tasks under *blocked conditions*—that is, multiple trials of one task (a block) are performed before moving on to the next task. Another group practices the same tasks under *random conditions*, with the same number of trials per task, but the order of performance of the tasks on successive trials is more or less random. Participants in the random-practice group are therefore continually switching from task to task, and they tend to perform more poorly than participants in the blocked condition, who spend a majority of the time

repeating the task from the previous trial. Despite this negative effect on acquisition performance, the random schedule ultimately results in better performance on delayed tests of retention and transfer. Note that studies of the effects of random practice differ from those of the variable–practice effect in that (a) learners are engaged in training on several distinct tasks, not multiple versions of a single task, and (b) the only factor that varies between groups in studies of blocked versus random practice is the order in which the tasks are practiced; the number of tasks practiced and the number of trials per task are identical for all learners.

The first clear demonstration of the random practice effect was presented in a paper by Shea and Morgan (1979). Three tasks were performed by each participant during the acquisition phase. Participants were to knock down three of six barriers with a tennis ball, in a specified order, as fast as possible. The specific set of three barriers to be knocked down was different for each task. Half of the participants practiced the three tasks under a blocked schedule, and the other half practiced them under a random schedule. Although performance of participants in the random group suffered relative to the blocked group during acquisition, they performed better on retention and transfer tests, regardless of whether those tests were administered with random or blocked schedules. (However, the advantage of random practice on the blocked retention test was not statistically significant.) Numerous studies on the random-practice effect were conducted subsequent to Shea and Morgan's study, confirming its robustness (see Magill & Hall, 1990, for a review).

Hall, Domingues, and Cavazos (1994) replicated Shea and Morgan's (1979) results and extended them, using a real-world sport (baseball) and skilled players as participants. They split a set of 30 collegiate baseball players into three groups, all of whom participated in normal batting practice. Two of the groups received two additional batting-practice sessions each week for 6 weeks, each consisting of 45 pitches—15 fastballs, 15 curveballs, and 15 change-up pitches. The third group, a control group, received no additional practice sessions. During the additional batting-practice sessions, one group (the blocked group) received 15 pitches of one type, then 15 pitches of the next type, followed by 15 pitches of the third type. The other group (the random group) received the same number of pitches of each type, but in a random order. At the end of the 6-week period, all 30 players received two 45-trial transfer tests, one in a blocked order and one in a random order.

Whereas during training, players in the blocked group produced a higher mean number of solid hits than did players in the random group, the opposite was true on the transfer tests. Players who trained under a random schedule performed significantly better at test, regardless of whether the test was administered under a blocked or random schedule.

Shea and Morgan (1979) and Hall and colleagues (1994) explored the effects of blocked versus random practice using motor skills. The effects of random practice are not, however, restricted to motor tasks. Carlson and Yaure (1990) tested the effects of blocked versus random practice on problem-solving efficiency and found results remarkably consistent with those found in the motor learning literature. Indeed, all of the practice effects discussed in the preceding sections—spaced practice, reduced feedback frequency, variable practice, and random practice—are quite general. They have been demonstrated using motor, verbal, and problem-solving tasks, diverse participant populations, and a wide range of retention intervals (for example, see Baddeley & Longman, 1978; Bahrick, 1979; and Dempster, 1990, on the spacing effect; see Salmoni et al., 1984, and Schooler & Anderson, 1990, on the effects of reduced feedback frequency; see Bird & Rikli, 1983; Gick & Holyoak, 1983; and Hunt, Parente, & Ellis, 1974, on variable practice; and see Carlson & Yaure, 1990; Landauer & Bjork, 1978; and Lee & Magill, 1983, on the random-practice effect).

Generality across domains and the tendency to produce opposite effects on short-term and long-term performance are not the only features shared by these practice manipulations. Note that the results of Hall and colleagues (1994) are in a general sense parallel to those observed by Kerr and Booth (1978) in their experiment on variable practice and to the results of Winstein and Schmidt (1990) in their study on reduced feedback frequency. In all three cases, performance at test was independent of the superficial similarities between practice and testing conditions. Recall that Kerr and Booth (1978) found an advantage of variable practice over constant (fixed) practice even when testing was conducted under formerly experienced fixed conditions. Winstein and Schmidt (1990) showed that practice with feedback after every trial produced poorer retention performance than practice with reduced feedback even when feedback was given after every trial at test. Also, Hall and colleagues (1994) demonstrated the superiority of random practice over blocked practice even when testing was blocked by task. Whereas intuitively one might expect that the ideal training con-

ditions would overlap maximally with testing conditions, these studies suggest that perhaps such superficial similarities are not as important as the underlying processes invoked at practice and the tendency of those processes to support long-term performance at test.

Interpreting Subjective Experience

Awareness of the limitations of objective performance measures is a crucial first step in assessing the effectiveness of training. However, equally important is a proper interpretation of subjective measures of training performance. There are at least two aspects of the training situation that can be assessed improperly—the performance of the learner and the effectiveness of the instructor. These two aspects can be viewed both from the perspective of the instructor and from the perspective of the learner, and the resulting subjective assessments can have a profound impact on the quality of training.

Interpreting the Learner's Performance

There are obviously two sides of an evaluation of training success: the instructor's and the learner's.

The Instructor's Perspective. Perhaps the most common misconception in the consideration of a learner's performance is that errors are always bad and successes are always good. To the contrary, inducing errors during training can be highly beneficial for learning. People *do* learn from their mistakes. In fact, one could argue that new learning only occurs after errors demonstrate the need for change. In research on concept identification, for example—in which participants are required to learn the rule or principle that determines whether a given multidimensional stimulus is or is not an instance of an experimenter-defined concept—there is clear evidence that learning only occurs after errors. As Trabasso and Bower (1968) put it,

> Opportunities for learning (entering the solution state from the presolution state) occur only in trials in which the participant makes an error, whereas correct response trials provide no opportunity to exit from the presolution state. (p. 46)

The broader point is that errors provide the stimulus for reassessment and discovery in a variety of learning contexts. Correct responding, which can often be the product of local cues—or of a conceptualization that works in the present situation, but not in general—offers no stimulus for change.

Individuals responsible for training can inadvertently structure

training such that trainees can respond correctly based on a superficial understanding, or, worse, so that constraints in the training situation shield trainees from exhibiting misunderstandings. Training conditions that keep performance levels artificially high—frequent feedback, massed practice, practice under fixed conditions—prevent learners from making mistakes, and are therefore likely to prevent significant learning.

Jacoby and colleagues (1994) warned that trainers who work to prevent errors during training may simply be deferring those errors to the posttraining environment, where errors can be costly. Yet even when trainers are aware of the value of mistakes during learning, management practices can serve to perpetuate substandard training practices. Trainers themselves are typically evaluated based on the performance of their trainees during training or at the end of training. This method of evaluation encourages trainers, consciously or unconsciously, to use training and testing methods that produce rapid improvement and high scores during training, which can, in turn, work in direct opposition to the long-term goals of training.

The Learner's Perspective. Bjork (1994b) and Jacoby and colleagues (1994) have argued that a learner's own subjective reading of the level of learning achieved is as important as the actual learning achieved. Whether trainees choose to engage in further practice or volunteer for other tasks depends on the reading they take of their own skills and knowledge. In certain work environments, such as air-traffic control, police operations, or the operation of nuclear power plants, individuals who do not possess critical skills and knowledge but think they do pose a special problem. In those environments, on-the-job learning can be hazardous not only to the individual, but to society as a whole. (For a thorough discussion of the importance of trainees' subjective experience during training and its effects on learning, see chapter 8, this volume.)

Learners, like their instructors, can also fall prey to an anti-error bias. As long as they are performing well and improving at a fast pace, they conclude that they must be learning; but when they make mistakes and improvement slows, they get discouraged and doubtful about the value of their training. Misunderstandings of this sort can lead learners to reject beneficial training programs in favor of less optimal instruction (Jacoby et al., 1994).

Baddeley and Longman (1978), for example, in an experiment that manipulated the practice schedules of British postal workers who

had to learn a new keyboard skill, observed a striking discrepancy between subjective preferences and objective measures. They varied the distribution of training sessions across time and found, consistent with earlier work, that distributing training sessions in time produced better learning per hour of instruction. At the end of the study, they asked the postal workers to rate how satisfied they were with their schedules. They found a negative relationship between those ratings and the actual efficiency of the different schedules. Left to their own judgment, the learners would have chosen the training condition that produced the smallest amount of learning per hour of instruction.

Misconstrual of the meaning of errors and successes is not the only type of misinterpretation demonstrated by learners when assessing their own level of learning. A related problem is that many people are not aware of the complex, multidimensional nature of human memory (Bjork, 1994b). Memory can be indexed in a variety of ways. Suppose we wish to test someone's memory for a particular episode—for example, exposure to a new song on the radio. There are several ways in which the testing can be carried out. One option is to simply ask: What was the new song you heard on the radio last week? Another method is to ask the person to choose from a list of alternatives. A more subtle approach is to find evidence of prior exposure by replaying the song and checking to see if the person can hum along. All three measures are valid options, but one cannot substitute for another.

A common error on the part of learners is to use one type of index to predict another (Bjork, 1994b; Jacoby et al., 1994). Most people can remember an occasion when, as a student, they walked into an exam highly confident of their mastery of the material and walked out equally sure that they had failed to master it. There is a good chance that a majority of our study time was spent rereading book chapters and class notes and nodding privately in agreement. If the exam had tested recognition of course material, that strategy might have been a successful one. On an exam that requires one to generate information from memory, however, such a strategy is ineffective; in effect, the exam requires a skill that may have been neglected during study. In other words, the feeling of familiarity experienced during study may have been inappropriately used to predict retrieval capability.

Experimental research conducted over the past 10 years has revealed numerous ways in which learners can use inappropriate indicators to judge their own knowledge. In many of these studies, researchers have been interested specifically in discovering whether learners' as-

sessments of their own knowledge could be affected by increased familiarity with certain types of information. One method of increasing a learner's feeling of familiarity for a piece of information is simply to expose the learner to that information. Prior exposure causes the information to become *activated*, or primed, in memory so that it is processed more easily if encountered again a short time later, and ease or fluency of processing is often experienced by the learner as a feeling of familiarity (see Jacoby, Kelley, Brown, & Jasechko, 1989). Thus, a common technique used by researchers has been to test the effects of prior exposure to information (and thus, familiarity for that information) on learners' judgments of their own related knowledge.

For instance, Reder (1987) conducted an experiment in which she used a game show paradigm to test the effects of prior exposure to parts of questions on feelings of knowing the answers to those questions. In her paradigm, participants quickly scanned a general-information question and indicated by pressing a button whether or not they thought they could answer the question. They were asked to base their response on a first impression rather than actually trying to retrieve the answer. After each trial, they attempted to answer the question and were subsequently shown the correct response. She showed that prior exposure to key words in a general-information question (such as the words *golf* and *par* in the question, "What is the term in golf for scoring one under par?") inflates the probability that participants feel they know the answer. It is interesting to note that such exposure to portions of the questions did not actually result in increased accuracy of the answers; the exposure only affected participants' *predictions* about their accuracy. In a similar vein, Reder and Ritter (1992) were able to influence participants' speeded judgments as to whether they could retrieve the answer to an arithmetic problem simply by exposing them to some of the terms of the problem. Given the problem "23 × 17," participants were more likely to judge the answer as retrievable if they had been previously presented the numbers 23 and 17 in the context of another mathematically distinct task. Thus, ease or fluency of processing of a question can sometimes be mistaken for knowledge of the answer.

Glenberg and Epstein (1987) demonstrated yet another form of faulty prediction on the part of learners. They had participants read text passages and rate their comprehension of the material. The participants were either experts or novices in the content domain of the passages (e.g., physics or music), but had no prior experience with the specific passages themselves. Later, all participants were asked to answer

questions about the passages. Experts were less accurate in assessing their comprehension than were novices; apparently, they were unable to separate their general familiarity with a domain from their comprehension of a specific piece of text.

It is interesting to note that even current ease of retrieval can be a misleading guide to judgments of a later capability to retrieve. Benjamin, Bjork, and Schwartz (1996) asked participants to answer each of a series of general-information questions and then to predict for each answer whether they would be able, at the end of the experiment, to recall that answer in the absence of the question. They found that participants' predictions of recall for a given answer correlated negatively with the time it initially took them to answer the question. Participants expected to recall an answer easily if they could answer the question quickly, and conversely, they expected to have difficulty with recall if answering the question required more time. In actuality, the probability of recalling an answer in the absence of cues was greater when more time was spent answering the original question, possibly because greater initial difficulty in arriving at an answer made for a more memorable episode. Participants' failure to appreciate this relationship reflects an underlying misconception about the degree to which performance on one memory retrieval task (recalling general information) can be used to predict performance on another (retrieving details of a particular episode).

Evaluating the Instructor's Effectiveness

Misinterpretations of the sort just described are not limited to inferences about the trainees' performance. The effectiveness of a trainer's instructional performance can also be misjudged. The consequences of incorrect appraisal can be serious, as perceptions of an instructor's effectiveness affect training in a number of ways. Trainers often use their own sense of success, or lack thereof, as a way of determining how much time to spend on different portions of the program. Errors in their judgment can lead to a suboptimal allocation of time and, therefore, suboptimal levels of learning. The perceptions of learners can also have a large impact on the training process. In many situations, evaluations of instructors by trainees are an important part of the review of training staff and practices. If those evaluations are misguided, they can result in changes that are detrimental to the training program.

The Instructor's Perspective. Research conducted by Newton (1990) bears on the issue of instructor effectiveness. Her studies are a powerful

demonstration of the effects of perspective on judgments of the comprehension of others. She separated participants into two groups—the "tappers" and the "listeners." The tappers' job was to choose a song from a list of 25 common titles and to tap the rhythm of that song for a listener. The listeners' task was to guess the name of the song. In addition, before the listeners responded, the tappers were asked to predict the likelihood that the listeners would identify the song correctly. The tappers estimated a 50% probability that listeners could identify the songs. In actuality, the listeners were correct only 2.5% of the time! Here, it is important to realize that the tappers probably heard much more in their own heads than the simple rhythms they were producing. Coversely, they may have heard a full rendition, complete with melody and harmony. Griffin and Ross (1991) suggested that the tappers were unable to adjust their estimates to take into account the differences between their perspective and the perspective of the listeners.

Instructors can easily suffer the same problems of perspective. Their expertise in an area of instruction can color their judgment of the clarity with which they communicate information. Extensive experience with training materials can also cause trainers to misjudge the difficulty of learners' tasks. Goranson (1976) provided people with a series of puzzles and asked them to judge the difficulty of the puzzles for others. Half of the people were given the answers from the outset and were asked to generate a judgment by pretending to solve the puzzle step by step. The other half were not given the answers and actually solved the puzzles themselves. The people who were given the answers to the puzzles grossly underestimated the time it would take to solve the problems as compared to those who actually solved them. Exposure to the answers subjected people to a hindsight bias (Fischoff, 1975) that prevented them from generating accurate assessments.

In another set of studies, Jacoby and Kelley (1987) asked participants to rate how difficult anagrams (such as FSCAR) would be for other participants to solve. Some of the anagrams were shown with the solution (FSCAR = SCARF), and others were shown without the solution (FSCAR = ?????). In the latter condition, participants' judgments of difficulty were considerably more accurate in predicting others' performance than in the former condition. Participants apparently used their own subjective experience (i.e., the ease with which they solved the anagrams) as an index for generating predictions about others. Exposure to the solutions robbed them of that subjective experience and therefore adversely affected their perceptions.

Jacoby and Kelley (1987) conducted another experiment using the same procedures, but instead of presenting some of the solutions beside the anagrams, they presented a list of half of the solutions to participants before they attempted to solve the anagrams. Exposure to the list made the solution of those items easier, but participants failed to take that prior experience into account in making their predictions. They judged anagrams for which they had seen the solutions as easier than those for which they had not. In short, they continued to base judgments of difficulty on their own subjective experience and therefore underestimated the difficulty of some items for others.

The Learner's Perspective. Learners can also be fooled when judging a trainer's effectiveness. They often mistake good presentation style for good teaching. Some providers of training seminars capitalize on this misperception. They put together an entertaining package—an instructor who resembles a stand-up comedian, a few cute illustrations, and some amusing activities—and thereby trade a little information for a lot of money. In a similar way, students in a classroom usually prefer an entertaining teacher to an effective, nonentertaining one. Of course, learners are probably more receptive to a trainer who makes things easy to understand, but if the content is lacking, training amounts to nothing more than a "feel good" session.

In a different vein, Jacoby and colleagues (1994) noted that learners can misattribute their own improvement for an improvement in instruction. As they accumulate knowledge and skill, they may find suggestions more useful and lectures easier to understand. However, instead of attributing their ease of understanding to their own level of proficiency, they might assume that the instructor is more organized or better prepared. Misattributions of this sort can open the door to abusive training practices. If the difficulty of the training materials is manipulated so that later tests are constructed to be easier than early ones, learners can be misled into believing that their improved performance reflects effective instruction.

In-Training Assessment Reconsidered

The multitude of misconceptions identified in the preceding sections may make the task of in-training assessment seem daunting. However, we offer some possible innovations that could make assessment during training a more realistic option. As evidenced by the findings reported

previously, research on the general principles of learning and memory, together with investigations within specific domains, constitute a resource to trainers in all types of organizations and industries. Research can aid in the design of training programs and in the interpretation of observed performance.

Making Subjective Experience More Diagnostic

Two of the major sources of difficulty in the previous assessment involve the use of subjective experience as a basis for assessment. Both trainers and learners often rely on subjective experience—rather than more objective measures—in forming opinions about the quality of training and in predicting future performance. (See Jacoby et al., 1994, for a more extensive treatment of this topic.) In general, the use of subjective experience in forming judgments is probably a good heuristic—especially in the absence of good objective methods of judgment (see, e.g., Wilson & Schooler, 1991). But as we have seen, subjective experience can sometimes be misleading in ways that can result in dramatic errors in assessment. However, it is fortunate that recent research suggests that training conditions can be structured in ways that better educate the subjective experience of learners and instructors.

Educating the Learner's Subjective Experience

In a previous section, we illustrated a number of ways in which learners sometimes base judgments of their comprehension or proficiency on inappropriate indices. For example, a learner might mistakenly use feelings of familiarity to predict performance on a test of recall. This phenomenon can be considered a misuse of subjective experience. Learners confuse their feelings of familiarity with an ability to recall information.

There are at least two ways to adjust for this effect. One is to educate learners about the nature of human learning, making them aware of the differences between recall, recognition, familiarity, and so on. Even when this is a viable option, however, it may not prevent learners from misjudging their own level of knowledge. Learners may not be able to adjust sufficiently for their subjective experience. For example, Strack, Schwarz, Bless, Kübler, and Wänke (1993) have shown that informing participants that they have been primed with information that will influence their assessments does little more than bias those later assessments in a direction opposite to that expected by the priming. Thus, it seems that such a proviso makes participants aware of

their ruined subjective experience but provides them no way to correct for it.

A second way to adjust for the deleterious effects of misleading subjective experience is to use objective measures of learning periodically during the course of training (Fischoff, 1975). Frequent tests that challenge the learners' understanding can serve to recalibrate their judgments of their own skill level. When faced with poor performance on a reliable test, learners will be much more likely to take steps to upgrade their skills or knowledge. Testing during training is, in fact, doubly advantageous. Tests themselves can be potent learning events. The very act of retrieving information from memory makes that information more recallable in the future (see, e.g., Anderson, Bjork, & Bjork, 1994; Bjork, 1975, 1988), and there is considerable evidence that a successful retrieval can be more advantageous than an additional study opportunity, particularly at a long retention interval (see, e.g., Allen, Mahler, & Estes, 1969; Hagman, 1983; Hogan & Kintsch, 1971; Landauer & Bjork, 1978).

Educating the Instructor's Subjective Experience

The steps that can be taken to educate an instructor's subjective experience are similar to those for the learner. The difference lies in the fact that the learners' attempts at assessment are directed toward themselves, whereas instructors attempt to gauge the learning of others. Consider the anagram study conducted by Jacoby and Kelley (1987). The task of participants was to rate the difficulty of the anagrams for others, and participants' ability to do so was negatively influenced by prior exposure to the solutions. In follow-up experiments, participants were informed of the effects of prior exposure to solutions before making their judgments, but they were still unable to compensate for their ruined subjective experience. The only manipulation that resulted in some adjustment of ratings was one in which participants were informed of the effect and asked to attempt to recognize the solutions as having been on the list before making their judgments.

Given instructors' difficulties in adopting the perspective of the learners, it may be wiser for instructors as well to rely on more objective measures of learning. However, one serious problem arises with the use of tests during the acquisition of a skill. Those tests can be extremely unreliable measures of long-term learning. This point was made at the outset, in the discussion of the distinction between learning and performance. There are several examples across domains of conditions of

practice that produce high levels of performance during training but yield poor retention and transfer performance at a delay (see, e.g., Schmidt & Bjork, 1992). For accurate assessment during training, tests need to be devised that are both (a) sensitive to the type of learning and comprehension that supports long-term performance and (b) relatively free of the fleeting influences of short-term factors.

Making Objective Performance More Diagnostic

The task of assessing long-term learning given performance on tests administered in the short term is clearly a daunting one, but it can be simplified by using measures in those assessments that are more diagnostic of the desired aspects of stable, long-term learning. What follows are some general characteristics to be considered when evaluating measures of performance on this dimension.

Assessing Transfer-Appropriate Processing

One general guideline for the selection of reliable tests of long-term learning rests on the concept of transfer-appropriate processing (Morris, Bransford, & Franks, 1977). Within the transfer-appropriate processing framework, a training manipulation is assumed to enhance retention or transfer to the extent that the processes exercised during training overlap with those required at retention. An implication of the foregoing statement is that instructors should employ tasks that require the learner to use the processes that will later be required to perform well in the target situation. Training regimens that encourage practice under fixed conditions—or that involve blocking skills by subtasks—work precisely against such a goal. However, practice that is carried out under variable conditions and schedules that require quasi-random switching from subtask to subtask are probably more similar to what trainees will encounter on the job and therefore are more desirable conditions for training.

In previous sections, we noted that typical measures of training performance do not always reflect long-term learning. If they did, conditions producing superior training scores would always produce superior retention scores. As noted previously, exactly the opposite is true of manipulations such as blocked versus random practice and full versus reduced feedback frequency. In pursuit of the goal of measuring learning during training, one solution is obvious: Use measures sensitive primarily to the factors contributing to enhanced retention.

In studies of blocked versus random practice, a common interpre-

tation as to why blocked practice appears good is that participants can bypass some memory retrieval processes in producing repetitions. In order to reveal the true level of competence being achieved, a trainer periodically could administer trials on a probe task that requires efficient retrieval of a training task, perhaps by embedding one of the old tasks in a more complex new task. In teaching kids to play basketball, for example, a coach could interrupt training drills periodically in order to assess whether his or her players have learned to perform different types of shots under random, rather than blocked, conditions. A test that simulated time pressure, varied positions of receiving the ball prior to shooting, varied defender positions, and so forth, would not allow the players to forgo those aspects of learning to retrieve and initiate the shot that are not exercised under blocked conditions.

This prediction implies yet another interesting prediction. By the logic just presented, the acquisition performance of participants in the random group is already reflective of processes supporting long-term retention. It is the performance of the blocked group that is potentially misleading. A similar conclusion applies to interpretations of the spacing effect. Because measurements of training performance for a spaced practice group are always taken after longer intertrial intervals, they are more similar to measures of retention than are measures of training performance for a massed practice group. Thus it seems that conditions that are better for learning are more likely to be the cases in which performance reflects learning. Put differently, those conditions of practice that enhance learning should also yield more accurate measures of learning.

Tests of Transfer During Training

Most tests administered during training can be considered tests of retention. They usually require the learner to reproduce the same information or skills that were experienced during instruction. Because retention tests consist of the same tasks practiced during training, their results are subject to the effects of the temporary influences alluded to earlier. Tests of retention are therefore almost always administered after a delay, when the temporary influences have dissipated. Transfer tests, by contrast, require learners to somehow transfer what they have learned to a novel task or altered conditions. Learners may be asked to draw inferences based on their knowledge or to perform a more complex version of a given task. Transfer tests composed of tasks similar to the training tasks are considered tests of near transfer, whereas tests

using very different tasks constitute tests of far transfer (for discussions of similarity and transfer, see Gick & Holyoak, 1987; Mayer & Greeno, 1972; Osgood, 1949).

Unlike the case for retention tests, it is generally acceptable for transfer tests to be administered immediately after the acquisition of skills. Exceptions would be those cases in which temporary factors present during acquisition, such as fatigue, might be expected to affect transfer tasks as well. In the typical cases in which this is not an issue, it also should be possible to administer transfer tests as probe tasks during training. The transfer tests would presumably provide purer measures of learning. And if, as Christina and Bjork (1991) have suggested, retention tests can be thought of as tests of very near transfer, then the probe tasks should be predictive of later retention performance.

To illustrate, consider a training task consisting of the repair of a certain type of machinery. Suppose further that the piece of machinery exists in several different sizes. In training, the instructor has a limited time to teach a somewhat complex repair skill and is faced with the choice of using machines of several sizes or of only one size throughout training. Realizing that the process of switching to new versions of a task almost always causes an increase in errors, the instructor decides to use a single size only. What the instructor does not realize is that the increase in errors during practice with several versions of a task is only a temporary performance decrement. Indeed, variable practice of this sort tends to enhance the flexibility of learners in performing a skill and allows greater generalization to new conditions.

If, during training, learners were given short tests requiring them to anticipate the effects of a change in the size of the machine—for example, difficulties in reaching certain parts or changes in the amount of force to be used in manipulating components—the advantage of variable practice might be revealed immediately. Learners engaging in variable practice would probably have more elaborate and flexible mental representations with which to approach the new task. Practice under fixed conditions would likely result in inferior performance on such a test. In other words, transfer tests like the one suggested previously would give trainers a much better sense of the capabilities of their trainees than would conventional tests of training performance. In this particular example, typical methods of assessment during training would have led the instructor to prefer a nonoptimal training plan.

The aforementioned methods of testing during training are, in

fact, likely to be triply advantageous. In addition to allowing more reliable assessment, they should potentiate future learning and induce desirable scheduling conditions during practice. As noted previously, successful attempts at retrieval tend to make information more recallable in the future (see Bjork, 1975, 1988) and can retard forgetting (Izawa, 1970). The effects of tests on practice scheduling should also be positive. By inserting short tests between training sessions, trainers will induce spaced and random practice—two conditions of practice that produce well-documented retention benefits (see Dempster, 1990; Magill & Hall, 1990). Thus it seems that testing during training, if carefully engineered, can have multiple beneficial effects on learning and evaluation.

Conclusion

We noted at the outset that, in general, the most desirable measure of training effectiveness is performance in the posttraining environment. The higher fidelity of tests in the real world coupled with the lack of short-term contaminating influences makes posttraining assessment a top choice in the evaluation of training. It is unfortunate, however, that posttraining assessment is very often out of the reach of those responsible for training. For this reason, our focus in this chapter has been on exploring the possibility of measuring effectiveness during training itself.

In identifying the numerous pitfalls one can encounter in assessing training performance—misleading performance, problems of perspective, misinterpretations of subjective experience, and so forth—we arrive at the conclusion that intuition and standard practice are often poor guides to the training process. The capability to evaluate training programs in progress rests in large part on educating ourselves about the nature of learning and its implications for performance. We have reached a point at which research findings provide the potential for significant improvements in the assessment of training.

We have suggested procedures that might upgrade the extent to which subjective experience is a valid measure of training, and we have suggested objective measures that might be more reliable indicators of long-term learning. Should those and related innovations prove viable as guides for selecting optimal configurations of training, in-training

evaluation, with its practical advantages over posttraining evaluation, may prove to be the assessment method of choice.

References

Alderman, R. (1965). Influence of local fatigue on speed and accuracy in motor learning. *Research Quarterly, 36,* 131–140.

Allen, G. A., Mahler, W. A., & Estes, W. K. (1969). Effects of recall tests on long-term retention of paired associates. *Journal of Verbal Learning and Verbal Behavior, 8,* 463–470.

Anderson, M. C., Bjork, R. A., & Bjork, E. L. (1994). Remembering can cause forgetting: Retrieval dynamics in long-term memory. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 20,* 1063–1087.

Baddeley, A. D., & Longman, D. J. A. (1978). The influence of length and frequency of training session on the rate of learning to type. *Ergonomics, 21,* 627–635.

Bahrick, H. P. (1979). Maintenance of knowledge: Questions about memory we forgot to ask. *Journal of Experimental Psychology: General, 108,* 296–308.

Benjamin, A. S., Bjork, R. A., & Schwartz, B. L. (1996). *The mismeasure of memory: When retrieval fluency is misleading as a metamnemonic index.* Manuscript submitted for publication.

Bjork, R. A. (1975). Retrieval as a memory modifier. In R. Solso (Ed.), *Information processing and cognition: The Loyola Symposium* (pp. 123–144). Hillsdale, NJ: Erlbaum.

Bjork, R. A. (1988). Retrieval practice and the maintenance of knowledge. In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory: Current research and issues, Vol. 1: Memory in everyday life* (pp. 396–401). New York: John Wiley.

Bjork, R. A. (1994a). Institutional impediments to effective training. In D. Druckman & R. A. Bjork (Eds.), *Learning, remembering, believing: Enhancing human performance.* Washington, DC: National Academy Press.

Bjork, R. A. (1994b). Memory and metamemory considerations in the training of human beings. In J. Metcalfe & A. P. Shimamura (Eds.), *Metacognition: Knowing about knowing* (pp. 185–205). Cambridge, MA: MIT Press.

Bird, A. M., & Rikli, R. (1983). Observational learning and practice variability. *Research Quarterly for Exercise and Sport, 54*(1), 1–4.

Carlson, R. A., & Yaure, R. G. (1990). Practice schedules and the use of component skills in problem solving. *Journal of Experimental Psychology: Learning, Memory, & Cognition, 16*(3), 484–496.

Carron, A. V. (1969). Performance and learning in a discrete motor task under massed vs. distributed practice. *Research Quarterly, 4,* 481–489.

Catalano, J. F., & Kleiner, B. M. (1984). Distant transfer in coincident timing as a function of variability of practice. *Perceptual & Motor Skills, 58,* 851–856.

Chamberlin, C., & Lee, T. D. (1993). Arranging practice conditions and designing instruction. In R. Singer, M. Murphy, & L. K. Tennant (Eds.), *Handbook of research on sport psychology* (pp. 213–241). New York: MacMillan.

Christina, R. W., & Bjork, R. A. (1991). Optimizing long-term retention and transfer. In D. Druckman & R. A. Bjork (Eds.), *In the mind's eye: Enhancing human performance.* Washington, DC: National Academy Press.

Cotten, D. J., Thomas, J. R., Spieth, W. R., & Biasiotto, J. (1972). Temporary fatigue effects in a gross motor skill. *Journal of Motor Behavior, 4,* 217–222.

Dempster, F. N. (1990). The spacing effect: A case study in the failure to apply the results of psychological research. *American Psychologist, 43,* 627–634.

Estes, W. K. (1955). Statistical theory of distributional phenomena in learning. *Psychological Review, 62,* 369–377.

Fischoff, B. (1975). Hindsight is not equal to foresight: The effects of outcome knowledge on judgment under uncertainty. *Journal of Experimental Psychology: Human Perception and Performance, 1,* 288–299.

Gick, M. L., & Holyoak, K. J. (1983). Schema induction and analogical transfer. *Cognitive Psychology, 15,* 1–38.

Gick, M. L., & Holyoak, K. J. (1987). The cognitive basis of knowledge transfer. In S. M. Cormier & J. D. Hagman (Eds.), *Transfer of learning: Contemporary research and applications* (pp. 9–46). San Diego, CA: Academic Press.

Glenberg, A. M., & Epstein, W. (1987). Inexpert calibration of comprehension. *Memory & Cognition, 15*(1), 84–93.

Goldstein, I. L. (1993). *Training in organizations: Needs assessment, development, and evaluation.* Pacific Grove, CA: Brooks/Cole.

Goranson, R. E. (1976). A paradox in educational communication. In I. Kusyszyn (Ed.), *Teaching and learning process seminars* (Vol. 1, pp. 63–76). Toronto, Ontario: York University Press.

Griffin, D. W., & Ross, L. (1991). Subjective construal, social inference, and human misunderstanding. In M. Zanna (Ed.), *Advances in experimental social psychology* (Vol. 24, pp. 319–359). New York: Academic Press.

Hagman, J. D. (1983). Presentation- and test-trial effects on acquisition and retention of distance and location. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9,* 334–345.

Hall, K. G., Domingues, D. A., & Cavazos, R. (1994). Contextual interference effects with skilled baseball players. *Perceptual and Motor Skills, 78,* 835–841.

Hogan, R. M., & Kintsch, W. (1971). Differential effects of study and test trials on long-term retention and recall. *Journal of Verbal Learning and Verbal Behavior, 10,* 562–567.

Hunt, R. R., Parente, F. J., & Ellis, H. C. (1974). Transfer of coding strategies in free recall with constant and varied input. *Journal of Experimental Psychology, 103*(4), 619–624.

Izawa, C. (1970). Optimal potentiating effects and forgetting-prevention effects of tests in paired-associate learning. *Journal of Experimental Psychology, 83,* 340–344.

Jacoby, L. L. (1978). On interpreting the effects of repetition: Solving a problem versus remembering a solution. *Journal of Verbal Learning and Verbal Behavior, 17,* 649–667.

Jacoby, L. L., Bjork, R. A., & Kelley, C. M. (1994). Illusions of comprehension, competence, and remembering. In D. Druckman & R. A. Bjork (Eds.), *Learning, remembering, believing: Enhancing human performance.* Washington, DC: National Academy Press.

Jacoby, L. L., & Kelley, C. M. (1987). Unconscious influences of memory for a prior event. *Personality and Social Psychology Bulletin, 13,* 314–336.

Jacoby, L. L., Kelley, C. M., Brown, J., & Jasechko, J. (1989). Becoming famous overnight: Limits on the ability to avoid unconscious influences of the past. *Journal of Personality and Social Psychology, 56,* 326–338.

Kerr, R., & Booth, B. (1978). Specific and varied practice of motor skill. *Perceptual & Motor Skills, 46,* 395–401.

Landauer, T. K., & Bjork, R. A. (1978). Optimum rehearsal patterns and name learning.

In M. M. Gruneberg, P. E. Morris, & R. N. Sykes (Eds.), *Practical aspects of memory* (pp. 625–632). London: Academic Press.

Lee, T. D., & Magill, R. A. (1983). The locus of contextual interference in motor skill acquisition. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 9*(4), 730–746.

Lee, T. D., Magill, R. A., & Weeks, D. J. (1985). Influence of practice schedule on testing schema theory predictions in adults. *Journal of Motor Behavior, 17,* 283–299.

Magill, R. A., & Hall, K. G. (1990). A review of the contextual interference effect in motor skill acquisition. *Human Movement Science, 9,* 241–289.

Mayer, R. E., & Greeno, J. G. (1972). Structural differences between learning outcomes produced by different instructional methods. *Journal of Educational Psychology, 63,* 165–173.

Melton, A. W. (1970). The situation with respect to the spacing of repetitions and memory. *Journal of Verbal Learning and Verbal Behavior, 9,* 596–606.

Morris, C. D., Bransford, J. D., & Franks, J. J. (1977). Levels of processing versus transfer appropriate processing. *Journal of Verbal Learning and Verbal Behavior, 16*(5), 519–533.

Newton, L. (1990). *Overconfidence in the communication of intent: Heard and unheard melodies.* Unpublished doctoral dissertation, Department of Psychology, Stanford University, Stanford, CA.

Osgood, C. E. (1949). The similarity paradox in human learning: A resolution. *Psychological Review, 56,* 132–143.

Reder, L. M. (1987). Selection strategies in question answering. *Cognitive Psychology, 19,* 90–138.

Reder, L. M., & Ritter, F. E. (1992). What determines initial feeling of knowing? Familiarity with question terms, not with the answer. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 18,* 435–451.

Salmoni, A. W., Schmidt, R. A., & Walter, C. B. (1984). Knowledge of results and motor learning: A review and critical reappraisal. *Psychological Bulletin, 95*(3), 355–386.

Schmidt, R. A. (1969). Performance and learning a gross muscular skill under conditions of artificially induced fatigue. *Research Quarterly, 40*(1), 185–191.

Schmidt, R. A. (1988). *Motor control and learning: A behavioral emphasis* (2nd ed.). Champaign, IL: Human Kinetics.

Schmidt, R. A., & Bjork, R. A. (1992). New conceptualizations of practice: Common principles in three paradigms suggest new concepts for training. *Psychological Science, 3*(4), 207–217.

Schooler, L. J., & Anderson, J. R. (1990). The disruptive potential of immediate feedback. *The Twelfth Annual Conference of the Cognitive Science Society* (pp. 702–708). Hillsdale, NJ: Erlbaum.

Shea, J. B., & Morgan, R.L. (1979). Contextual interference effects on the acquisition, retention, and transfer of a motor skill. *Journal of Experimental Psychology: Human Learning and Memory, 5*(2), 179–187.

Strack, F., Schwarz, N., Bless, H., Kübler, A., & Wänke, M. (1993). Awareness of the influence as a determinant of assimilation versus contrast. *European Journal of Social Psychology, 23,* 53–62.

Trabasso, T., & Bower, G. H. (1968). *Attention in learning: Theory and Research.* New York: Wiley.

Van Rossum, J. H. (1990). Schmidt's schema theory: The empirical base of the variability of practice hypothesis: A critical analysis. *Human Movement Science, 9*(3–5), 387–435.

Wilson, T. D., & Schooler, J. W. (1991). Thinking too much: Introspection can reduce the quality of preferences and decisions. *Journal of Personality and Social Psychology, 60*, 181–192.

Winstein, C. J., & Schmidt, R. A. (1990). Reduced frequency of knowledge of results enhances motor skill learning. *Journal of Experimental Psychology: Learning, Memory, and Cognition, 16*(4), 677–691.