Performance Measurement in Simulation-Based Training: A Review and Best Practices
Eduardo Salas, Michael A. Rosen, Janet D. Held and Johnny J. Weissmuller
Simulation Gaming 2009; 40; 328 originally published online Dec 22, 2008;
DOI: 10.1177/1046878108326734

The online version of this article can be found at:
http://sag.sagepub.com/cgi/content/abstract/40/3/328

Published by:
SAGE
http://www.sagepublications.com

On behalf of:
Association for Business Simulation & Experiential Learning
International Simulation & Gaming Association
Japan Association of Simulation & Gaming
North American Simulation & Gaming Association
Society for Intercultural Education, Training, & Research

Additional services and information for Simulation & Gaming can be found at:

Email Alerts: http://sag.sagepub.com/cgi/alerts
Subscriptions: http://sag.sagepub.com/subscriptions
Reprints: http://www.sagepub.com/journalsReprints.nav
Permissions: http://www.sagepub.com/journalsPermissions.nav
Citations http://sag.sagepub.com/cgi/content/refs/40/3/328
Simulation-based training (SBT) is a methodology for providing systematic and structured learning experiences. The effectiveness of this methodology is dependent on the quality of performance measurement practices in place. Performance measurement during SBT must be diagnosed; that is, the causes of effective and ineffective performance must be determined. This diagnostic measurement drives the systematic decisions concerning corrective feedback and remediation. The purpose of this article is to provide a state of the science review of human performance measurement systems in SBT. To this end, three specific goals are addressed. First, a review of the theoretical foundations being used to drive performance measurement systems in SBT is provided. Second, an overview of the methodologies and approaches to measurement in SBT is provided. Third, a set of best practices for designing performance measurement systems for use in SBT are provided. These best practices are based on the scientific and practice-based literatures.

Keywords: causes; corrective feedback; design; diagnosis; diagnostic measurement; effectiveness; feedback; methodologies; performance; performance measurement; performance diagnosis; best practices; remediation; simulation-based training; SBT; teams; team performance; theoretical foundations

In domains such as the military, medicine, business, and aviation, the present-day workplace is characterized by high degrees of complexity and competitiveness. In this environment, maximizing human performance is essential for safety, effectiveness, organizational development, and even survival. Human performance can be improved through effective training programs. Performance measurement is a key component of these programs, allowing trainers to identify areas of weakness and to provide targeted feedback and remediation.

Authors’ Note: We would like to thank Jean MacMillan, George M. Alliger, and David Crookall for comments on an earlier version of this article. This work was supported by the Navy Personnel Research, Studies and Technology Department and the Bureau of Naval Personnel (Selection & Classification Office) under the auspices of the U.S. Army Research Office Scientific Services Program administered by Battelle (Delivery Order 0665, Contract No. DAAD19-02-D-0001).
resilient, adaptive, and flexible in ambiguous and information-intensive contexts, but it can also be plagued with error and inefficiencies. Preparing people for performance in complex environments requires a complex approach to training. Exclusive use of the more traditional methods for training personnel (e.g., lecture-based training programs) fails to meet the demands of many work environments found in modern organizations, and consequently, organizations increasingly turn to simulation-based training (SBT) methods. The value of SBT primarily stems from its ability to provide practice opportunities in environments that replicate important features of the “real world” environment. However, to be maximally effective, SBT must provide guided practice to ensure the correct competencies are acquired. Without guidance, no learning or even negative learning—learning of ineffective performance processes—may occur. The feedback provided to trainees about their performance constitutes a major source of guidance in SBT, and feedback is generated from an understanding of trainee performance in a simulation scenario. Therefore, the measurement of performance underlies the capacity of SBT to prepare personnel (or evaluate their readiness) for complex environments and for organizations to manage, improve, and sustain high levels of effectiveness. Simply put, if performance is not explicitly defined and measured, it cannot be changed or improved in a systematic way. This is especially true in SBT, where many goals, purposes, and reasons drive the need to measure performance. These include, but are not limited to, supporting personnel selection, diagnosis, feedback in the training process, assessment of training outcomes, assessment of interventions, and development.

Generally, performance measurement can be viewed as an investment in time and monetary expense that pays dividends in information on which decisions can be made and action can be taken (Brannick & Prince, 1997). Whether dealing with team or individual performance, the ability to measure and diagnose performance accurately is necessary to meet any of the previously mentioned goals. Consequently, a suboptimal approach to performance measurement not only squanders the time and other resources required to implement a performance measurement system but also may incur additional costs engendered by poor decision making and improper actions made on the basis of data derived from poor performance measurement practices. This is the primary motivation of this article: to provide scientifically rooted and practically relevant guidance for developing performance measurement systems in SBT that maximize the return on investments made in performance measurement.

Therefore, this article sets out to answer a series of questions regarding the state of the art and science in human performance measurement systems in SBT. First, what are the theories driving performance measurement research and practice? What are the theoretical foundations behind the development of performance metrics and measurement systems capturing performance during SBT? These questions are salient because (a) measurement should always be grounded in theory and (b) theory can provide some very practical guidance for developing performance measures. Therefore, we present a review of theories of individual and team performance relevant
to SBT. Second, what methodologies are being used? Theory helps answer the question of what to measure; answering the question of how to measure it is the next logical step. In doing so, it is important to examine a wide range of measurement techniques and tools available for use today. Our review discusses what types of performance measurement tools are used and how, when coupled with a theoretical understanding of performance, they can be used to increase learning during SBT. Third, what are the “best practices” in current performance measurement systems? The best practices presented in this article are based in the discussion of theory and methods as well as a review of the practice-based literature. In SBT, constraints (e.g., limited training time, limited number of trainers/observers) frequently interfere with implementing an ideal plan for performance measurement or training in general. Examples from the practice-based literature illustrate how these constraints are managed in the application of best practices. Answering the previously stated questions necessitates examining the requirements of performance measurement systems and assessing the degree to which a performance measurement technique or system fulfills these requirements.

**What Theories Are Driving Performance Measurement in SBT?**

When developing a performance measurement system, defining explicitly what to measure can often be one of the most difficult tasks. The more complex the performance to be trained, the more difficult this becomes. As simulations often represent highly complex tasks, designers of SBT curricula must face this challenge successfully to produce effective training. Human performance is essentially behavior in the context of doing a task, but it is not always clear what behaviors are important components of performance. Theories of performance serve as valuable resources for resolving this confusion. They provide a description of the critical components of performance that must be captured and used to generate corrective feedback.

The purpose of the measurement system and the nature of the task being trained guide the selection of theories, which subsequently influences the choice of measures and methods of measurement (Cannon-Bowers et al., 1989). This implies that developing a performance measurement system can involve a broad repertoire of theories. For example, an SBT curriculum designed to train a task with a high psychomotor component (e.g., a laparoscopic surgery trainer) should draw on theories of motor learning and production to identify and capture critical aspects of performance. Similarly, an SBT curriculum designed to train Crew Resource Management skills in flight crews should draw on theories of teamwork and team performance to develop performance measures and generate feedback. These tasks vary in substantive ways. Consequently, designing simulation-based practice activities and accompanying measurement tools to develop effective performance in these tasks requires application of different theories of the underlying nature of performance.
Additionally, the purpose of the performance measurement system (i.e., what decisions will be made with the data) will influence the type of measures selected. The requirements for data used for the purposes of assessment differ from those for data used to generate formative feedback. These distinctions will be expanded on throughout this article.

This section reviews several of the major theories used in the development of team and individual performance measurement systems. As vast numbers of human performance theories exist, the review is selective rather than comprehensive. It is intended to highlight the types of theories available and the utility of incorporating theory into the design of performance measurement systems for SBT.

**Individual Performance**

Performance measurement on the individual level has been driven by numerous theories. This section provides an overview of some of the more influential perspectives, including frameworks of learning outcomes, general theories of human performance, human information processing (HIP), and theories of expertise. The major theories discussed in this section are summarized in Table 1 along with important implications for performance measurement in SBT.

**Frameworks of Learning Outcomes**

SBT, and training in general, should promote learning on the part of trainees. Learning outcomes are “persistent states that make possible a variety of human performance” (Gagné, 1984, p. 377). From a training perspective, these constitute the classes of things that must be acquired to perform the tasks targeted for training. However, learning and performance are highly task dependent. This is true for the learner (i.e., the ability to transfer learned knowledge and skill depends on the similarity of the practice and transfer tasks) as well as for the theories that describe learning and performance. That is, theories of motor performance do not provide much help for understanding a coordination or decision-making task. To provide a conceptual framework for organizing the diverse literature on learning and performance, researchers have developed categories of learning outcomes (e.g., Bloom, 1956; Gagné, 1984; Kraiger, Ford, & Salas, 1993). These categories represent classes of learning outcomes wherein theories of learning are generalizable within, but not across, categories. One such framework proposes three general categories: cognitive outcomes (i.e., declarative knowledge, knowledge structure, and cognitive strategies), skill-based outcomes (i.e., skill compilation, and automaticity), and affective outcomes (i.e., attitudes and motivation) (Kraiger et al., 1993). Each of these categories of learning outcomes indicates different implications for the nature of performance as well as the methods for measurement.
The outcomes of learning drive effective performance, and different types of tasks require different types of learning outcomes (cognitive, skill, and affective); theories of learning and performance are generalizable within, but not across, categories of learning outcomes.

Performance is determined by three primary factors: declarative knowledge, procedural knowledge and skill, and motivation. People perform various internal manipulations of information received from the environment; performance is determined by the efficacy of these transformations.

Superior performance is a function of specialized psychological mechanisms developed during extended practice activities. Superior performance is a function of specialized psychological mechanisms developed during extended practice activities.

Critical performance processes in cognitively intensive tasks are not directly observable. Measurement in SBT should capitalize on the ability to control events in the scenario to provide good opportunities to make inferences about covert aspects of performance. Critical performance processes in cognitively intensive tasks are not directly observable. Measurement in SBT should capitalize on the ability to control events in the scenario to provide good opportunities to make inferences about covert aspects of performance. Critical performance processes in cognitively intensive tasks are not directly observable. Measurement in SBT should capitalize on the ability to control events in the scenario to provide good opportunities to make inferences about covert aspects of performance. Critical performance processes in cognitively intensive tasks are not directly observable. Measurement in SBT should capitalize on the ability to control events in the scenario to provide good opportunities to make inferences about covert aspects of performance.

Different types of learning outcomes require different approaches to performance measurement. The development of performance measures in SBT should be guided by an analysis of the learning outcomes involved in the task being trained. Performance is determined by the degree to which trainees possess needed competencies (learning outcomes). Complex performance in simulations is determined by sets of different competencies. Performance measures must be developed to capture these competencies independently during practice activities to generate corrective feedback. Performance is determined by the degree to which trainees possess needed competencies (learning outcomes). Complex performance in simulations is determined by sets of different competencies. Performance measures must be developed to capture these competencies independently during practice activities to generate corrective feedback. Performance is determined by the degree to which trainees possess needed competencies (learning outcomes). Complex performance in simulations is determined by sets of different competencies. Performance measures must be developed to capture these competencies independently during practice activities to generate corrective feedback. Performance is determined by the degree to which trainees possess needed competencies (learning outcomes). Complex performance in simulations is determined by sets of different competencies. Performance measures must be developed to capture these competencies independently during practice activities to generate corrective feedback.
Table 1 (continued)

<table>
<thead>
<tr>
<th>Level</th>
<th>Theory</th>
<th>Description</th>
<th>Implications for PM in SBT</th>
</tr>
</thead>
<tbody>
<tr>
<td>Team</td>
<td>IPO models</td>
<td>Team performance is a function of various input, process, and output variables.</td>
<td>SBT is the primary means of capturing team processes.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>In addition to team processes, performance measurement systems must capture the relevant input variables that account for performance during practice activities.</td>
</tr>
<tr>
<td></td>
<td>Shared mental models</td>
<td>Team performance is facilitated by shared knowledge held between team members.</td>
<td>SBT scenarios can be developed to measure shared mental models (e.g., &quot;Do team members understand the needs of their fellow team members during a performance episode?&quot;).</td>
</tr>
<tr>
<td></td>
<td>Adaptability</td>
<td>Team performance over time involves adapting.</td>
<td>Team performance should be captured over time and leverage the dynamic nature of simulation.</td>
</tr>
<tr>
<td></td>
<td>Big five of teamwork</td>
<td>A prototype of team performance involves five factors (team leadership, mutual performance monitoring, back-up behavior, team orientation, and adaptability) and three coordinating mechanisms (shared mental models, closed loop communication, and mutual trust).</td>
<td>Measurement of team performance in SBT should capture changes in team performance in response to changes in the simulation scenario.</td>
</tr>
<tr>
<td></td>
<td>Macrocognition/shared cognition</td>
<td>Teams are information-processing units.</td>
<td>There are multiple aspects of teamwork that should be captured during SBT practice activities.</td>
</tr>
<tr>
<td></td>
<td></td>
<td></td>
<td>Performance measures can be developed for each of these factors and coordinating mechanisms to provide diagnostic measurement and corrective feedback for individual- and team-level competencies.</td>
</tr>
</tbody>
</table>

The flow of information within a team during SBT is a critical aspect of performance that should be captured.
As systematic training, SBT involves specifying learning objectives and competencies (i.e., the knowledge, skills, and attitudes that underlie effective performance). With a specification of what competencies are targeted for training in a given simulation, learning outcomes frameworks can be used to make better decisions about how to measure the targeted competencies. For example, consider the use of an automobile simulator to develop driving skills in novices. Initially, a targeted competency may be teaching the coordination of movements (e.g., wheel control, shifting, brake and acceleration) necessary to successfully drive the car. This learning objective would fall within the skill compilation category of learning outcomes suggesting that what must be measured and subsequently converted into feedback for the trainee are the relative motion patterns of the trainee. This information describes the degree to which the trainee has achieved smooth control. As the trainee progresses and masters these fundamental skills, cognitive strategies (e.g., changing manner of driving under differing weather conditions) may be a targeted competency. This indicates that the performance that must be captured involves adaptation of driving strategies as conditions demand (e.g., adjusting following distance in rain versus good driving conditions).

Theory of Performance and Performance Determinants

Campbell, McCloy, Oppler, and Sager (1993) have advanced a high-level theory describing the major components of performance across a wide range of jobs. This theory states that performance is a multidimensional construct comprising distinct components (e.g., job-specific task proficiency, written and oral communication, facilitating peer and team performance) and that relevant variance in these distinct performance components can be accounted for by three factors, or determinants of performance (Campbell et al., 1993). According to their model, performance is a function of (a) declarative knowledge, (b) procedural knowledge and skill, and (c) motivation. These categories are consistent with the learning outcomes categories previously discussed. Declarative knowledge is defined as the facts, principles, goals, and self-knowledge necessary to complete a given task and more broadly can be considered an understanding of the task requirements. Procedural knowledge and skill involves the cognitive, psychomotor, physical, self-management, and interpersonal skills relevant to a task; these skills are attained when knowing what to do (declarative knowledge) is combined with knowing how it is done. Campbell et al. (1993) define motivation as a result of three choice behaviors, including the choices concerning the expenditure of effort, the level of effort, and the persistence of effort. This theory has received direct empirical support, although the model tested assumed that the three performance determinants functioned in an additive manner, when a multiplicative relationship is more likely (McCloy, Campbell, & Cudeck, 1994). For example, an additive model suggests that acceptable levels of performance can be achieved with high levels of two determinants (e.g., declarative knowledge and
procedural knowledge and skill) and none of the third determinant (e.g., motivation). In a multiplicative model, there must be at least some minimal level of all three performance determinants for adequate levels of performance to be manifested. For a recent review of general theories of job performance, see Tubré, Arthur, and Bennett (2006).

Campbell’s model fits well with a learning outcomes view and contributes the idea that overall performance is a function of different levels of each of the learning outcomes; that is, for effective performance, trainees must have appropriate levels of the necessary learning outcomes. For SBT, this has the implication that, in most situations, multiple categories of learning are relevant and therefore multiple methods of measurement should be employed. Performance in the complex tasks frequently represented by simulations must be decomposed into its constituent elements and represented within the performance measurement system. This is critical in order to provide diagnostic feedback. For example, if a trainee fails in a simulation designed to build interpersonal skills for sales personnel, is that a function of knowledge (i.e., the trainee didn’t know what to do), skill (i.e., the trainee didn’t know how to apply knowledge about interpersonal skills), or motivation (i.e., the trainee didn’t want to exhibit good interpersonal skills)? These three options and any combination thereof are possible, and all require different feedback to correct the deficient competencies. It is very difficult to accomplish this without developing separate measures for each competency.

In the previous two sections, we reviewed general theories of performance. These are valuable tools for analyzing a task and designing measurement in SBT to increase learning; however, within each learning category (or determinant of performance), there are a variety of theories that can be drawn on. In the following section, the cognitive learning category is explored in more detail.

HIP

Much of the current understanding of human performance on the individual level is rooted in theories of HIP, especially for the large and increasing number of cognitive tasks that dominate current work settings. Numerous instances of HIP models of human performance exist (e.g., Card, Moran, & Newell, 1983), stemming from early discrete stage models (e.g., Broadbent, 1972; Posner, 1978; S. Sternberg, 1969). The essence of HIP models of human performance is that information in different forms flows from the environment or system (comprising people and technology) through the individual. The individual operates various types of transformations on this information; these transformations are the sources of effective and ineffective performance (Wickens & Carswell, 1997). Information-processing theories have evolved in the past decades, and a recent emergence has been the use of field studies to develop context-rich models of HIP in “real world” settings. The Naturalistic Decision Making (NDM) movement is representative of this approach (Klein, Orasanu,
Calderwood, & Zsambok, 1993). As one of the main goals of NDM is to create an understanding of real-world cognitive work, NDM theories and frameworks are suitable for forming a basis for developing performance measures. One widely used NDM model is the Recognition Primed Decision (RPD) model.

The RPD Model. The RPD Model (G. Klein, 1993) represents a different type of HIP model, one developed from an ecological perspective of HIP in natural settings outside the laboratory (Wickens & Carswell, 1997). This type of model fits in a class of macrocognitive models (Cacciabue & Hollnagel, 1995; G. Klein et al., 2003), which are characterized by a focus on cognitive work as it naturally occurs in situations involving complex decisions, time pressure, high risk, high stakes, ill-defined goals, and conditions where all key variables are not under the control of the individual. The RPD model describes an expert’s decision-making process as comprising two general subprocesses:

(a) cue and pattern recognition
(b) mental simulation of courses of action

First, a decision maker senses cues in the environment and attempts to match the pattern of cues in the decision environment with past experiences. If the match for the present problem is found, the decision maker retrieves the following:

- associated expectancies about the problem
- possible courses of action that are likely to be effective
- a set of cues that are most relevant to the problem
- goals for the situation

If a reasonable match is not immediately found, the expert decision maker engages in a more exhaustive diagnosis of the situation, which involves gathering more informational cues about the problem.

Once the situation has been assessed and relevant information retrieved from memories of similar past experiences, the expert decision maker mentally simulates a course of action to determine if the solution will work given the contextual constraints that characterize the current situation. The decision maker’s mental simulation results in either the adoption of the original course of action, modification of particular aspects of the retrieved course of action so that it is more effective in the present situation, or rejection of the course of action in lieu of retrieval and mental simulation of an alternate course of action. The RPD model makes explicit the crucial role of situation assessment and situation awareness (SA) in expert decision-making processes in natural real-world settings. Endsley (1997) proposes that the three levels of SA (i.e., perception of elements in the environment, comprehension of the current situation, and projection of future status) are the inputs to the decision-making process. That is,
processes culminating in an understanding of the current and likely future states of the environment feed directly into the decision-making process. This process is essential to the subsequent cue recognition and retrieval of expectancies, goals, relevant cue sets, and possible courses of action stages of the RPD model.

HIP models call to light a significant challenge for measuring performance; many critical aspects of performance are not observable. Covert and internal strategies and thought processes (i.e., the transformations people make on information) constitute the essence of performance in many cases, but these are not directly accessible to an observer. Information about the degree to which an individual possesses knowledge-based or cognitive strategies must be inferred through various means. In training, this is normally accomplished with various knowledge tests; however, in SBT, the scenario can provide a powerful means for making these distinctions. With a well-crafted scenario, the trainee can be required to respond to trigger events designed to elicit behavior tied to different components of covert performance (e.g., pattern recognition or mental simulation in RPD). This technique will be described in more detail in the section on event-based measurement.

Expertise

The study of human expertise—of reliably superior performance—has a long history in the scientific community (Ericsson, 2006). Traditionally, this research has taken place in task domains such as chess, sports, and music. However, expertise research increasingly makes contributions to the understanding of performance in complex tasks within organizations in two ways. First, theoretical work concerning the nature of expertise developed in tightly defined tasks has contributed to an understanding of performance in more ill-defined tasks (e.g., those usually trained with SBT; Salas & Rosen, in press). Second, more cognitively complex “real world” tasks are more frequently becoming the subject of expertise research (e.g., Sonnentag, Niessen, & Volmer, 2006). For example, the RPD model described earlier is rooted in the expertise literature and was first developed to describe the performance of firefighters.

Expertise is best understood as a set of psychological and physiological adaptations to the constraints of a task or domain of tasks (Ericsson & Lehmann, 1996). That is, the nature of expertise is different depending on the domain of expertise (Ericsson & Smith, 1991). The extensive empirical literature on expert performance across different domains has been synthesized to reveal a set of characteristics that contribute to expertise across domains to varying degrees depending on the nature of the task (see Bedard & Chi, 1992; Chi, 2006; Feltovich, Prietula, & Ericsson, 2006; Glaser, 1987; Glaser & Chi, 1988; Hoffman, Feltovich, & Ford, 1997). These characteristics, or mechanisms of expertise, are illustrated in Figure 1 and should be considered as a “prototype” of expertise (Sternberg, 1997), where each of these mechanisms may contribute to expert performance in a given domain depending on the task constraints. In the simplest of cases, expert performance can be viewed as a
function of extended experience within a domain and intentional practice activities that yield a larger, more organized knowledge base (Bordage & Zacks, 1984; Chi, Feltovich, & Glaser, 1981) and memory skill (Ericsson & Kintsch, 1995), which in turn facilitates pattern recognition and problem representation abilities (Glaser & Chi, 1988; Simon & Chase, 1973; Zeitz, 1997), as well as self-monitoring skills and automaticity (Lesgold et al., 1988; Moors & De Houwer, 2006). Each of these mechanisms requires practice to acquire; they can be represented in performance measurement systems, and performance during SBT can be used to generate corrective feedback and accelerate the learning process.

Increasingly, simulation is used not only as a means to develop expertise (i.e., through providing practice opportunities with diagnostic feedback, the primary mechanisms of building expertise) but also as a means to study expertise (Ericsson & Ward, 2007). Performance measurement is equally important in both of these situations; however, the approaches taken differ to a large degree. Seeking to understand expert performance within a simulation will likely involve rich qualitative approaches (e.g., protocol analysis, interview techniques) that identify which mechanisms are used and
the specifics of their application. Once the specific mechanisms of expert performance have been identified, these can guide the development of more quantitative techniques for capturing performance of developing experts within the same simulation. This is true for diverse types of performance and simulations such as that within a large-scale emergency response simulation or a marksmanship trainer.

Team Performance

Organizations increasingly use team-based work arrangements to manage the growing complexity of modern work. These team-based systems provide many advantages; they increase safety by providing barriers to errors; they provide a means to address problems too complex for any one individual; they provide a means for rapidly calling on diverse expertise. This section provides an extensive review of major theoretical approaches to understanding team performance and the factors that influence it. A review of this breadth is provided for three main reasons:

1. First, SBT is a core approach to developing teamwork skills. Team performance is dynamic and requires the interaction of numerous individuals responding to a changing task environment. Practice-based activities that afford this type of dynamic interaction are therefore key to developing teamwork skills.
2. Second, teamwork is generally more complex than individual performance. It involves not only multiple individuals’ performance task work but their coordinated action. Consequently, the process of developing performance measures can be more difficult and the need for good theory more salient.
3. Modern work, and the simulations that represent it, frequently involve more than one person.

Like all measurement, valid measures of team performance must be based on theory (Salas, Burke, Fowlkes, & Priest, 2003). However, there is little consensus on definitions of team performance and team effectiveness (Kendall & Salas, 2004; MacBryde & Mendibil, 2003), which in turn complicates the process of creating general measures of team performance and effectiveness that are firmly rooted in theory. The issues involved with defining team performance should not come as a surprise given the struggle to define performance on the individual level (Arvey & Murphy, 1998; Campbell et al., 1993; Motowildo, 2003). The plethora of models of teamwork can complicate the matter of choosing one to guide the development of team performance measures. Team performance measurement should be customized to the context and grounded in a careful consideration of what is appropriate for a particular team (Kendall & Salas, 2004; MacBryde & Mendibil, 2003). Targeted performance processes and outcomes must always be defined in a manner that is relevant to the situational factors of the team being studied. Factors to consider are team type, team goals and objectives, and assessment objectives (Kendall & Salas, 2004).
In this vein, frameworks for developing team performance measures suited to a particular context have been formulated (e.g., Cannon-Bowers & Salas, 1997; MacBryde & Mendibil, 2003; Smith-Jentsch, Johnston, & Payne, 1998). According to Smith-Jentsch et al.’s (1998) framework, there are four dimensions of team performance that should be addressed by a measurement system: team processes, individual processes, team outcomes, and individual outcomes. Individual-level measures are included in team measurement to identify and assist the team members that need help developing their particular task work skills. In addition to team outcomes, team processes are an essential component of a measurement system (Cannon-Bowers & Salas, 1997). Four major team processes that have been empirically related to performance outcomes and that should be considered for inclusion in the measurement system are information exchange, communication, supporting behavior, and team leadership (Smith-Jentsch et al., 1998). However, it is recognized that even though there are some processes that are likely essential to all team performance outcomes, certain teams operating in specific contexts will probably require the development of unique measures to capture idiosyncratic or distinctive team objectives and patterns of interaction (Kendall & Salas, 2004). Therefore, it is necessary to clearly define the team processes of interest in any one particular training context before attempting to formalize a measurement system.

SBT can be used to train a broad variety of teams. The aviation, military, and power generation industries have long-standing traditions of SBT for teams. Additionally, the U.S. health care system is moving toward widespread adoption of SBT for teams. However, all teams are not created equal. The specific team performance strategies targeted for training and measurement within a command and control simulation may be different from those trained and measured within a simulation designed to build teamwork in Emergency Department staff.

This section is dedicated to a review of the major conceptualizations of theoretical work with implications for the measurement of team processes. Specifically, the topics of Input, Process, Output (IPO) models, shared mental models, team adaptability, the “big five” of teamwork, and macrocognition/team cognition are discussed.

**IPO Models**

The number of theories and frameworks of team effectiveness and performance is vast—each theory or framework representing different research threads contributing to the larger picture of understanding teams (Salas, Stagl, & Burke, 2004). However, from this large and diverse field of team research emerges the central conceptual theme of current team research—the IPO model. Models and theories of this class seek to characterize the relationship between input variables, such as individual and team characteristics; process variables, such as shared cognition, leadership, communication, coordination, decision making, and back-up behavior; and outcome variables, such as performance outcomes, productivity, and satisfaction (e.g., Gersick, 1988).
By illuminating the relationships between IPO variables, these models and theories characterize teamwork as dynamic and multidimensional as well as highlight the importance of team processes in effective teamwork (Guzzo & Dickson, 1996).

IPO models and theories in the literature have been systematically evaluated and an integrative framework has been produced (Tannenbaum, Beard, & Salas, 1992). This framework, illustrated in Figure 2, identifies four classes of input variables (i.e., work structure, individual characteristics, team characteristics, and task characteristics). The input variables relate to each other as well as to the team processes occurring over time. Team processes influence performance outcomes, which are in turn looped back to team input variables in the form of feedback. Explicit in the Tannenbaum et al. (1992) framework is the importance of training in moderating the relationships between input and process variables as well as the relationship between process and outcome variables. SBT provides opportunities for team members to engage in the dynamic process behaviors and receive feedback on their performance.
Shared Mental Models

Definitions of mental models vary, but generally, they can be thought of as knowledge structures involved in the integration of information and the comprehension of a given phenomenon (Johnson-Laird, 1983). Subsequently, a mental model that is shared is a representation of information held by more than one person—team members. A shared mental model is an organized knowledge structure that facilitates interactions between team members (Klimoski & Mohammed, 1994). This “sharedness” of information models facilitates the effective coordination of action as well as promotes a similar method for processing new information within the team. These knowledge structures shared by individuals support the team members’ development of accurate causal explanations of their environment and tasks as well as accurate projections of future states of the environment and tasks. Consequently, shared mental models allow the team members to better adapt to the environment comprising task demands and their fellow team members (Cannon-Bowers, Salas, & Converse, 1993). This increased ability gained from shared mental models to better diagnose the environment and project future states has been linked with better decision making in teams (Stout, Cannon-Bowers, & Salas, 1996), in particular their ability to recognize causal relationships in the external environment, make sound inferences, and create better explanations (Cannon-Bowers et al., 1993; Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000).

The boon of shared mental models is especially significant in conditions that complicate communication between team members (e.g., extremely high workload, physical distribution, time pressure, stress). An increasing characteristic of teamwork is that team members are distributed in space and rely on various communication technologies to interact with each other. In this situation, shared mental models become crucial because they allow team members to predict the needs (e.g., information, material resources, present workload) of their teammates (Mathieu et al., 2000). Shared mental models account for fluid, adaptive, and implicit coordination often observed in highly effective teams (Mohammed & Dumville, 2001).

Adaptability

Historically, there has been a paucity of research that considered the temporal dynamics of team processes and performance (see Dyer, 1984), resulting in a lack of formalized understanding of how teams perform over time. However, this omission is being remedied by more recent models (e.g., Gersick, 1988; Morgan, Salas, & Glickman, 2001). Adhering to a multidisciplinary, multiphasic, multilevel perspective grounded in theory from cognitive, human factors, and industrial/organizational psychology, Burke, Stagl, Salas, Pierce, and Kendall (2006) recently proposed a theory of team adaptation. At the center of team adaptation is adaptive team performance—defined as
an emergent phenomenon which compiles over time from the unfolding of a recursive cycle whereby one or more team members utilize their resources to functionally change current cognitive or behavioral goal directed action or structures to meet expected or unexpected demands. (p. 1190)

The team adaptation model is illustrated in Figure 3.

Burke et al. (2006) used an IPO approach to delineate a series of phases that produce adaptive team performance. Phase 1 consists of situation assessment comprising a cue recognition process and an ascription of meaning to cue patterns in the environment. Phase 2 is characterized by plan formulation, the phase in which the team develops and decides on a course of action. Phase 3 is plan execution, which relies on team coordination mechanisms. The final phase consisting of team learning, Phase 4, arises from an evaluation of the team’s performance. Knowledge of this evaluation will change how the team carries out the earlier phases of the model on future passes through the adaptive cycle. Therefore, as teams progress through these phases, emergent affective and cognitive states develop, which are then called on by team members in future performance episodes. In this way, Burke et al. characterize team adaptation as an episodic and recursive process.
This theory of team adaptation has significant ramifications for performance measurement in simulation. In a manner similar to traditional team performance, adaptive team performance can be viewed as a property of the team or as a configurational construct (Kozlowski, Gully, Nason, & Smith, 1999). Therefore, depending on the organizational context, task flow, interdependencies, and other contextual variables, adaptive team performance will emerge in different ways (K. Klein & Kozlowski, 2000). Burke et al. (2006) propose that the ideal way to capture adaptive team performance is through a patterned emergence model. To capture adaptive team performance, metrics must be dynamic (i.e., capture performance over time) as well as context sensitive (i.e., they must be able to represent and be responsive to the changing nature of the operational environment).

The “Big Five” of Teamwork

The “big five” of teamwork, illustrated in Figure 4, is an effort to unify the fragmented literature on teamwork and team effectiveness into a manageable set of features that can be used in practical contexts (Salas, Sims, & Burke, 2005). Salas et al. (2005) propose that the “big five” components of teamwork—leadership, adaptability, mutual performance monitoring, back up behavior, and team orientation—are important in all team contexts. In addition to these five factors, three coordination mechanisms—shared mental models, closed loop communication, and trust—are necessary for effective teamwork. Shared mental models and adaptability have been reviewed earlier; the remaining components of the “big five” of teamwork are discussed subsequently.

Leadership. Research shows and we know intuitively that leaders have a considerable impact on team and organizational effectiveness (e.g., Zaccaro, Rittman, & Marks, 2001). Increasingly, a functional perspective to leadership has been adopted (Fleishman et al., 1991; Hackman, 2002; Zaccaro et al., 2001), a perspective rooted in and understanding of leadership as

social problem solving that promotes coordinated, adaptive team performance by facilitating goal definition and attainment. (Salas, Burke, & Stagl, 2004, p. 343)

Functional leadership can be thought of in terms of four types of behavior that leaders engage in as they respond to social problems:

(a) information search and structuring
(b) information use in problem solving
(c) managing personnel resources
(d) managing material resources

Increasing complexity in the task and social environments demands increased levels of adaptive reactions by the team and the team leader. Recently, the concept of
Figure 4
The Big Five Model of Teamwork

Source: Adapted from Salas, Sims et al. (2005).
shared leadership has been explored as a means to increase the adaptive nature of team leadership and thereby increase team performance.

In contrast to traditional vertical leadership with more formal and static roles, *shared leadership* can be defined as “the transference of the leadership function among team members in order to take advantage of member strengths (e.g., knowledge, skills, attitudes, perspectives, contacts, and time available) as dictated by either environmental demands or the development stage of the team” (Burke, Fiore, & Salas, 2004, p. 105). Shared leadership has been shown to be more effective than more traditional forms of leadership (Pearce & Sims, 2002). This increased performance has been attributed to the fact that shared leadership enables the team members to shift leadership functions among one another in a manner that affords maximal exploitation of individual-level knowledge and skills. Formal hierarchical leadership is not replaced by shared leadership; it is in fact the formal leadership that creates the team climate and structure that support the sharing of leadership within a team. The degree to which leadership functions can be relocated within a team—a coordination process—determines the effectiveness of this type of leadership. This is an area of particular relevance to SBT. Because of the dynamic nature of shared leadership (team members must shift leadership functions as task demands change), dynamic practice is a vital component of training and evaluation.

*Team orientation.* Team orientation has been defined in a way that extends past a mere preference for working with teammates; it includes the tendency to coordinate, evaluate, and use the task inputs of other teammates and patterns of behavior that improve team and individual performance (Driskell & Salas, 1992). Empirical research findings have shown that when people experience stress their attention narrows as they focus inward; Individuals become less willing to accept feedback/input from others as they tightly focus their efforts on their individual task (Driskell & Salas, 1991; G. Kleinman & Serfaty, 1989). Feedback and input from others is critical for obtaining optimal levels of performance in many situations and vital to a team’s ability to learn and adapt within complex environments as well. By providing motivation and a higher baseline tendency to work as a team, team orientation may partially protect against the effects of stress within teams. Decision-making performance is also positively influence by team orientation (Driskell & Salas, 1992); this is an essential skill for teams that need to adapt their performance to changing circumstances.

*Mutual performance monitoring.* Mutual performance monitoring is the ability to “keep track of fellow team member’s work while carrying out their own . . . to ensure that everything is running as expected and . . . to ensure that they are following procedures correctly” (McIntyre & Salas, 1995, p. 23). Mutual performance monitoring can have negative effects on performance if it is perceived as a means of keeping tabs on team members or team members trying to protect themselves from responsibility for errors. That is why team attitudes must be accepting of mutual performance
monitoring, and the team culture must make provisions for practice as a valid way of elevating levels of team performance.

In addition to appropriate team attitudes, certain cognitive factors in the team are required for effective mutual performance monitoring. Specifically, team members must share mental models of the task, team, and equipment to correctly recognize abnormalities in their fellow team members’ performance as well as the needs of their teammates while engaged in mutual performance monitoring. Essentially then, the team members must know what should be happening before observing what actually is happening, which can yield fruitful information and boost individual and team performance.

Back-up behavior. Mutual performance monitoring is a necessary but not sufficient condition for developing and maintaining high levels of team performance. To capitalize on the mutual performance monitoring behaviors, team members must engage in back-up behavior. Back-up behavior is “the discretionary provision of resources and task-related effort to another . . . (when) there is recognition by potential back-up providers that there is a workload distribution problem in their team” (Porter et al., 2003, pp. 391-392). Back-up behavior is essential for effective team performance in three fundamental ways:

(a) it affords the opportunity for team members to provide feedback to one another to improve performance
(b) it allows team members to provide assistance to one another during task performance
(c) it allows for an adaptive capacity in teams such that the team can readjust its strategies by dynamically shifting workload between team members when an overload is detected in any one specific team member (Marks, Zaccaro, & Mathieu, 2000).

Back-up behavior takes the form of either verbal or physical aid.

Coordinating mechanisms. The five core components of teamwork function by means of three core coordination mechanisms: shared mental models (discussed earlier), closed loop communication, and mutual trust. Communication can generally be defined as the “exchange of information between a sender and a receiver” (McIntyre & Salas, 1995, p. 25). Communication processes bind the team’s work together and has been described as the glue of teamwork. Needed information is distributed throughout the team by means of communication. Similarly, shared cognitive structures are developed and their accuracy and completeness maintained through communication; this is necessary for creating common ground and adaptive performance.

Research findings show that effective teams adapt their communication strategies by alternating between implicit and explicit communication. When a high-performing team engages in explicit communication, that team will most frequently engage in a particular type of explicit communication: closed loop communication. Closed loop communication is defined by the following three dimensions:

Downloaded from http://sag.sagepub.com at PURDUE UNIV LIBRARY TSS on May 23, 2009
(a) a message being initiated by the sender
(b) the message being received, interpreted, and acknowledged by the intended receiver
(c) a follow-up by the sender ensuring that the message was received and appropriately interpreted (McIntyre & Salas, 1995)

The closed loop communication process is the most effective method of explicit communication because it ensures that there was no distortion of the message by the communication process—that is, closed loop communication confirms that the message was interpreted by the receiver in the manner intended by the sender.

**Mutual trust.** In addition to shared mental models and closed loop communication, mutual trust is an essential coordinating mechanism according to the “big five” of teamwork. *Mutual trust* has been defined in team contexts as “the shared perception . . . that individuals in the team will perform particular actions important to its members and . . . will recognize and protect the rights and interests of all the team members engaged in their joint endeavor” (Weber, 2002, p. 205). Without mutual trust, team members will waste attentional and communication resources checking to ensure that other team members are performing their tasks appropriately (R. Cooper & Sawaf, 1996). With mutual trust, team members are free to spend those resources of time and effort collaborating and working on individual tasks. Additionally, trust is implicated in a number of team processes and outcomes, such as willingness to disseminate information (Jones & George, 1998), member participation and contribution, outcome quality, and member retention (Bandow, 2001).

**Macrocognition/Shared Cognition**

Just as macrocognitive aspects of individual HIP have been emphasized in recent research, the growing tendency to conceptualize groups and teams as information processing systems (Hinsz, Tindale, & Vollrath, 1997) is accompanied by an investigation of team-level macrocognitive processes and team cognition in real-world settings. The rise in importance of teams for addressing complex work environments is paralleled by an increase in the amount of cognitive tasks allocated to teams. For example, teams are formed to assess situations, make decisions, design and monitor plans, and gather knowledge. Therefore, understanding and measuring cognition at the team level is a vital aspect of understanding team performance and effectiveness. However, lack of clear definitions of the constructs involved with team cognition, the measurement purpose, and the influence of the task environment have often been cited as impediments to the development of appropriate measurement methods for team cognition (e.g., Mohammed, Klimoski, & Rentsch, 2000). This section begins with a general discussion of the state of team cognition measurement, specifically issues revolving around the level of analysis and aggregation of individual data to team-level representations. This section concludes with a more detailed discussion...
of several classes of knowledge states and cognitive processes generally classified under the heading of shared or team cognition.

Team cognition can be characterized as an explanation of the interaction of and dependencies between intraindividual- and interindividual-level processes (Fiore & Schooler, 2004). Therefore, team cognition is a multilevel construct, and measurement methodologies must address this aspect of the nature of team cognition to create valid and reliable measures. In the bulk of the research to date, there are three categories of measures used to assess team cognition: elicitation methods, team metric methods, and aggregation methods (Cooke, Salas, Cannon-Bowers, & Stout, 2000). All three of these methods share the underlying assumption that team knowledge is a collection of individual-level knowledge. The three categories of techniques also build on one another such that each successive category is designed to allow for an extrapolation of individual-level data to data representative of team-level cognition.

First, the knowledge elicitation approach focuses on extracting a representation of domain-specific, mental model–like knowledge, usually from an expert in the domain. This is usually accomplished with such techniques as observation, interview, mapping techniques, and protocol analysis. In general, the knowledge elicitation category of team cognition measurement has been applied to relatively long-lasting and stable knowledge structures at the individual level. That is, these types of techniques assess mental models that are assumed to change over time but at a relatively low rate. These techniques have not been widely applied to more dynamic, rapidly changing knowledge structures such as team situation models. However, there are notable exceptions to this generalization. Specifically, knowledge elicitation techniques have been applied to the assessment of team situational assessment (Cooke, Stout, & Salas, 1997, 2001).

The second category of measures is team metrics. This category of measures arises to fill the need of quantifying information gathered through the knowledge elicitation methods. Much of the data derived from knowledge elicitation techniques are in the form of graphical representations and therefore difficult to analyze in several dimensions:

1. in terms of accuracy of the knowledge structure
2. comparisons between individual team members (e.g., similarity or overlap measures)
3. aggregations of individual-level data to the team level

Metrics can be calculated from the knowledge elicitation data in terms of similarity (e.g., a percentage score for overlap in individual knowledge structures), or accuracy (e.g., a percentage score for correct knowledge), as well as metrics that account for team heterogeneity and distribution of knowledge.

Lastly, team metrics need to be effectively translated into team-level measures of cognition. This is done through various aggregation methods, such as averaging the individual scores or taking the median, maximum, or minimum value. There is no
clear method to aggregate individual-level data to team-level data in this manner, as in some instances taking the mean value of a team metric will be misleading, particularly when the variance in individual scores is high. Cooke et al. (2000) recommend using individual metrics to categorize teams versus using an aggregate team measure. In this way, relationships between various configurations of teams can be explored, a capability not easily realized with aggregate data.

Researchers are currently exploring options to measurement techniques that attempt to aggregate individual-level data to team-level representations of cognition, as some conceptualizations of team cognition imply that the aggregation-based methods are not tapping appropriate types of knowledge and cognition. For example, Klimoski and Mohammed (1994) describe team knowledge as an emergent phenomena resulting from the interaction of team members. This perspective places a greater emphasis on team processes and allows for the possibility that team knowledge is something different than the sum of the team’s individual members’ knowledge. Cooke, Salas, Kiekel, and Bell (2004) present a framework of team knowledge that elucidates the realm of knowledge (i.e., individual and relatively static structures) tapped by the traditional methods described above as well as the team knowledge that is left untapped (i.e., knowledge involving team processes such as communication, SA, and coordination). The rest of this section discusses several shared cognition constructs in more detail: transactive memory, team SA, and metacognition.

Transactive memory. The construct of transactive memory is an expansion on most conceptualizations of Shared Mental Model (SMM) in that it speaks to the storage of information, not the interrelationships among the stored items. Wegner (1986) defines a transactive memory system as (a) a combination of the knowledge possessed by each individual and (b) a collective awareness of who knows what. Yoo and Kanawattanachai (2001) argue that it is team members’ metaknowledge about who knows what in a team. Austin (2003) expands this definition by adding transactive memory consensus/agreement and accuracy as components. The transactive memory system is founded on the idea that, because of complexity, individuals know only part of what the team as a whole knows and this team knowledge is distributed unequally among members (Moreland, Argote, & Krishnan, 1996). Research has begun to indicate that transactive memory systems can enhance a team’s performance, particularly when the task is complex and requires a considerable contribution of knowledge from individual team members (Faraj & Sproull, 2000; Lewis, 2000; Moreland, 1999).

Team situation awareness (TSA). Situation awareness (SA) has been defined as the “reception of the elements in the environment within a volume of time and space, the comprehension of their meaning, and the projection of their status in the near future” (Endsley, 1995, p. 36). SA varies on many dimensions, but accuracy and completeness are of greatest interest. The accuracy of an individual’s SA is the degree to which a component of knowledge in an individual’s SA represents the
The completeness of an individual’s SA is the degree to which knowledge in individual’s SA captures all or only some of critical environmental factors. As represented in the above definition, there are three levels of SA: perception, comprehension, and projection. Accuracy and completeness apply to each level of SA.

Although often described at an individual level of analysis, researchers have begun to conceptualize TSA. For example, Salas, Prince, Baker, and Shrestha (1995) suggest that TSA is more than the sum of the individual members’ SA and includes team process behaviors as well. Stout et al. (1996) argue that the work on TSA has a number of implications:

1. existing mental models shape individual SA
2. shared mental models seem to be critical to team performance because they allow members to form adequate explications and expectations of task and team actions
3. TSA enhances team performance

After conducting critical incidents reviews with pilots, Prince and Salas (1997) identified four major actions important for TSA in commercial pilots: identifying problems or potential problems, demonstrating knowledge of the actions of others, keeping up with flight details, and verbalizing actions and intentions.

Metacognition. Flavell (1979) conceptualized metacognition as deliberate, goal-directed, and future-orientated mental behaviors that are directed toward accomplishing a task. Kluwe (1987) extended this work and identified two attributes of metacognitive activity. First, the thinking subject has some knowledge of his or her own thinking and the thinking of others. Second, the attribute consists of active monitoring and regulation of one’s own thinking, where one is acting as the causal agent of his or her own thinking. Consistent with Kluwe’s work, Mazzoni and Nelson (1998) describe two primary functions of the metacognitive system: control and monitoring. Metacognitive control refers to activities taken by individuals to influence the course of their own thinking, including initiating, monitoring, and prioritizing mental activities. Metacognitive monitoring consists of updating the metalevel model of the situation with information from the object level. This model can then be evaluated with respect to the goal the system is attempting to achieve.

Research investigating team members’ higher order cognitive structures has focused on illuminating the role of metacognition (Cohen, Freeman, & Wolf, 1996; Kozlowski, 1998; Lord & Emrich, 2001; Sternberg, 1998). It has been asserted that “at the team level, metacognitive and self regulatory processes must extend beyond the self” (Kozlowski, 1998, p. 120). This line of research suggests that teams can become self-learning, adaptive systems when they have developed metacognitive skills at the team level. Accumulating research findings suggest that team member metacognition can be accurately measured and that obtained scores are related to
important team outcomes such as leader performance (Koslowski, Gully, Smith, Nason, & Brown, 1995; Marshall-Mies et al., 2000), team performance (Marsh, Kiechel, Boyce, & Zaccaro, 2001), and tacit knowledge and social competencies (Banks, Bader, Fleming, Zaccaro, & Barber, 2001).

In sum, a large array of theoretical frameworks, models, and approaches have been used to conceptualize various aspects of human performance at the individual and team levels. This is not surprising given that human performance in general is a multifaceted and highly complex process. A complex approach to measurement must be taken to accurately capture a complex phenomenon. Multiple measurements rooted in the various theoretical approaches discussed above can help to provide a deep and detailed view of human performance in training environments. To this end, measurement methodologies are discussed in the next section.

What Methods Are Being Used?

The theories outlined in the previous section are valuable tools for understanding performance in complex simulations. To maximize outcomes of SBT, these critical aspects of performance must be captured during practice activities, and corrective feedback must be generated. There are a variety of methods in use to accomplish this goal at the individual and team levels. These range from quantitative to qualitative and from subjective (e.g., self-report measures) to objective (e.g., task outcomes). A third dimension by which to categorize performance measures is the level at which they capture performance (e.g., individual or team). This aspect of performance measurement is less solid than the issue of whether the measure is qualitative or quantitative, subjective or objective. With due care, some of the same methodologies can and have been used to measure different levels of performance. However, the nature of the construct greatly affects how that construct is measured in a multilevel manner. Two main “pure-type” distinctions involved in developing measures of team-level performance based on individual performance are compilation and composition. Composition describes a process whereby a lower level (e.g., individual) emerges upward to the team level but remains essentially the same. Conversely, compilation is the process whereby similar but distinctively different lower level properties combine into a higher level (e.g., team) property that is related to but different from its diverse lower level constituent parts (Kozlowski & Klein, 2000).

Another important factor in performance measures is the implicit conceptualization of performance that the measure embodies; that is, does the measure originate from the perspective that performance is a process or an outcome? Campbell (1990) argues that “performance is not the consequence or result of action; it is the action itself” (p. 174). This sentiment is echoed in the team research, specifically, that all too often team performance is conceptualized as the products or outcomes of actions that extend over time and not the actual processes themselves (Kozlowski & Bell,
However, this approach does not dismiss the importance of performance outcomes, as they are often what is ultimately of interest. Essentially, if performance is conceptualized and measured as a process and this aids in the performance measurement goals of diagnosis and intervention, then there is the underlying assumption that by improving process, outcomes will also benefit.

The rest of this section is dedicated to an overview of some of the performance measures in use. Measures are grouped into two major classes—qualitative and quantitative. Other aspects of the measure such as level of analysis and specifics about what the measure captures are discussed for each measure. Table 2 summarizes the measures discussed in this section. It details the type of data the measure relies on, the level of analysis, whether the measure captures processes or outcomes, and key sources.

**Qualitative Methods**

Many of the techniques classified as qualitative are used to develop an understanding of what performance consists of in a given domain. That is, qualitative measurement techniques can be used in the development of performance measurement systems, which are subsequently used to capture performance during SBT. Qualitative measurement techniques are best suited for building an understanding of the processes that drive effective performance in the real world (e.g., the goal of SBT). More quantitative measures of performance can then be developed that capture these processes and enable feedback, remediation and assessment during SBT.

**Protocol analysis.** In support of measuring internal individual and team processes, protocol analysis is a method for using verbal reports as indicators of cognition (Ericsson & Simon, 1993). This is an increasingly valuable tool for training and simulation as the tasks trained become more complex and cognitively based. There are two general categories of protocol analysis: on-line and off-line (Shadbolt, 2005). On-line variants involve the participant verbalizing his or her thoughts concurrently with task performance. In off-line protocol analysis techniques, the participant views a video recording of his or her task performance session and verbalizes his or her thoughts. Generally, on-line analysis techniques are thought of as more naturalistic, whereas off-line techniques are assumed to interfere less with task performance. An additional related technique is the cognitive walkthrough, which involves participants verbalizing how they would perform a task but without actually doing the task (Bainbridge & Sanderson, 2005).

Protocol analysis is a three-stage process leading from research or design questions to conclusions (Bainbridge & Sanderson, 2005). First, the data are collected; that is, the verbalizations are recorded and transcribed. At this stage, it is important to capture as much of the task context as possible so that ambiguities arising in the analysis at a later time can be resolved. For this reason, video recording is preferred to just audio.
### Table 2
**Summary of Methodologies**

<table>
<thead>
<tr>
<th>Type of Data/Analysis</th>
<th>Method</th>
<th>Level of Analysis</th>
<th>Example Applications in SBT</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Qualitative</strong></td>
<td>Protocol Analysis</td>
<td>X</td>
<td>In a business simulation: used to capture the internal strategies of an expert negotiator; this can then be used to generate learning objectives and performance measures for less expert negotiators</td>
</tr>
<tr>
<td></td>
<td>Critical Incident Technique</td>
<td>X</td>
<td>Used to elicit critical events that can be used in the design of scenarios to elicit behaviors tied to competencies targeted for training</td>
</tr>
<tr>
<td></td>
<td>Concept Mapping</td>
<td>X</td>
<td>Used to elicit required domain knowledge and structure of expert knowledge</td>
</tr>
<tr>
<td></td>
<td>Behavioral Observation Scales (BOS)</td>
<td>X</td>
<td>In a naval fleet operations simulation: observers can rate the quality of communication between commanding officers</td>
</tr>
<tr>
<td></td>
<td>Communication analysis</td>
<td>X</td>
<td>In an aviation simulation designed to develop Crew Resource Management skills: the degree to which flight crews that use “closed loop” communication can be coded</td>
</tr>
<tr>
<td></td>
<td>Event-based measurement</td>
<td>X</td>
<td>In health care team training: a team member can be “overloaded” with patient issues and the degree to which other team members who respond with back-up behavior can be measured</td>
</tr>
<tr>
<td></td>
<td>Structural knowledge assessment</td>
<td>X</td>
<td>In a strategic planning simulation: the degree to which an individual’s knowledge structure has changed and approximates an expert knowledge structure can be assessed and used to drive feedback and remediation</td>
</tr>
<tr>
<td></td>
<td>Self-report measures</td>
<td>X</td>
<td>In an interpersonal skills simulation: the trainee can self-report levels of motivation to exhibit the behavior targeted for training</td>
</tr>
<tr>
<td></td>
<td>Automated Performance Recording</td>
<td>X</td>
<td>In a golf trainer: characteristics of the motion patterns of a swing can be captured and turned into visual feedback and quantitative metrics.</td>
</tr>
</tbody>
</table>
It is also good practice to have the participant review the transcripts if possible. Second, the protocol data are prepared for analysis. The researcher identifies a general structure of related segments in the protocol. This resembles a task analysis in that the utterances are grouped into hierarchical sections of activity. Next, the analyst identifies meaningful units that appear to represent separate mental processes. From this decomposition of the protocol, a structure of mental activities can be inferred, formal descriptive languages can be developed to aid in statistical analysis and abstraction of the protocol, and inferences about processes not represented in the protocol can be made (Bainbridge & Sanderson, 2005). The third and final stage of verbal protocol analysis is the examination of explicit and implicit content achieved through various methods of content analysis, sequential analysis, and modeling techniques.

**Critical incident technique.** Since its development in the Aviation Psychology Program of the U.S. Air Force during World War II, the Critical Incident Technique has been applied in a diverse range of research domains (see the American Psychiatric Association Critical Incident Technique Bibliography, Fivar, 1980). Flanagan (1954) describes the critical incident technique as a “set of procedures for collecting direct observations of human behavior in such a way as to facilitate their potential usefulness in solving practical problems” (p. 327). An incident is sufficiently complete human behavior that, in and of itself, allows for inferences about that agent. An incident is critical if it occurs in a situation in which (a) the intent of the agent is sufficiently clear to the observer and (b) there is little doubt about the effects of the behavior, as the observer can clearly identify the consequences of the action. The technique does not compose a hard and fast set of immutable rules but a flexible set of principles that must be adapted to specific situational characteristics. There are three core characteristics of the technique: Observers are only required to make the simplest of judgments, only qualified observers participate, and an a priori agreed-upon purpose of the research is used to guide the evaluation of observations. These elements of the Critical Incident Technique serve to minimize the amount of observer bias and interpretations in the observations. The technique can be used to measure typical performance and proficiency.

Additionally, the Critical Incident Technique is often used to develop scenarios and behavioral triggers embedded in scenario-based training and therefore a valuable tool in developing event-based measurement systems (e.g., Fowlkes, Lane, Salas, Franz, & Oser, 1994; Morgan, Lassiter, & Salas, 1989). A more recent adaptation of the Critical Incident Technique is the Critical Decision Method, which is concerned with determining factors influencing situation assessment and decision making (Klein, Calderwood, & MacGreggor, 1989). In many ways, the Critical Decision Method is a hybrid between the Critical Incident Technique and retrospective protocol analysis in that it focuses on probed interviews with individuals concerning performance on nonroutine events. The Critical Decision Method is also related to Cognitive Task Analysis (Hoffman, Crandall, & Shadbolt, 1998).
Concept mapping. An important technique in measuring team knowledge, a central component of team cognition (Cooke et al., 2004), is concept mapping or cognitive mapping. This knowledge elicitation approach measures the structure, content, and robustness of the knowledge of individuals and teams and generates a graphical representation of how cognitive concepts are related to one another for any one person. This technique derives much of its popular support from its treatment and analysis of meaning as a relation between knowledge elements. Because these discrete elements of cognitive content can be related in numerous ways (e.g., causality, resemblance, continuity, proximity, etc.; Weick & Bougon, 1986), many different types of concept maps exist; however, maps representing causality between elements are generally the most widely used (Mohammed et al., 2000), as causality is considered the most important relation between knowledge elements (Gray, Bougon, & Donnellon, 1985).

In addition to the existence of different types of concept maps based on the relations represented in them, there are several different methods for generating the maps. First, data are gathered from the participants in one of two general approaches: (a) questionnaires and interviews (Mohammed et al., 2000) or (b) through analysis of preexisting data (Carley, 1997). Second, in relation to the questionnaire and interview techniques, there are several variations based on the whether the content, structure, or both or neither are provided to the participants (Salas, Priest et al., 2005). In one instance, the research gives the participant a set of concepts, which must be placed within a predefined structure. Another case involves the participant being given only the structure and then required to generate the content on his or her own. Lastly, participants may be required to generate both the structure and the content.

Quantitative Methods

The measurement approaches discussed earlier are primarily used to develop an understanding of the criterion; they help to answer the “what to measure” question. After this understanding of what drives performance outcomes in a domain, quantitative measures that sample this criterion space can be developed. This section discusses the major types of quantitative measurement tools available for capturing performance during SBT.

Behaviorally anchored rating scales (BARS). BARS are a widely used technique first developed by Smith and Kendal (1963) and a very popular way of promoting accurate and consistent observer ratings when rating and classifying individual and team behaviors. BARS consist of a series of brief descriptions that characterize different levels of performance. Each description is associated with a specific scale value, and the behavioral examples represent episodes of excellent, acceptable, and poor performance. These descriptions of behaviors and anchor points are frequently generated from Subject Matter Expert (SME) accounts using the Critical Incident Technique (Flanagan, 1954; Fowlkes et al., 1994; Morgan et al., 1989), which can be used to characterize the upper (i.e., high-quality performance) and the lower (i.e.,
poor performance) bounds or anchor points of the BARS. Because BARS include
descriptions of specific behaviors, one concern in the use of BARS involves the ten-
dency for observers to focus only on the behaviors that are a part of the checklist
(Murphy & Constans, 1987; Murphy & Pardaffy, 1989).

**Behavioral observation scales (BOS).** Like BARS, the purpose of BOS is to facil-
itate accurate observer ratings of team performance; however, BOS assess typical
performance, unlike BARS, which assesses isolated performance. BOS accomplish
this through the use of a Likert-type scale that observers use to record the frequency
with which teams exhibit certain behaviors and engage in targeted processes. The
team behaviors and processes must first be conceptually defined and examples gen-
gerated from SMEs in a manner akin to that used in BARS. These examples are then
distilled into cogent and concise statements. During the simulation or training exer-
cise, the observer then issues a numerical frequency score to each of the process or
behavior statements (e.g., ranging from 0 = never to 9 = always). Because this
method of observational rating forces the raters to evaluate over a period of time and
rely on their memories, it is important that observers be trained to avoid falling prey
to known human memory contaminants to data such as primacy and recency effects
(e.g., Baddeley & Hitch, 1993). Meta-analytic review has shown that it is possible to
achieve and maintain high reliability and accuracy of frequency-based judgments as
ratings despite some inherent biases (e.g., overestimating infrequent events and
underestimating highly frequent events).

**Communication analysis.** As of now, team processes are best measured through com-
munication analysis. Most frequently, communication has been measured in terms of
simple frequency, as this approach is the most straightforward and easy to implement.
However, research attempting to link frequency of communication with team outcomes
has produced mixed results (e.g., Foushee & Manos, 1981; Jentsch, Sellin-Wolters,
Bowers, & Salas, 1995; Orasanu, 1990; Orasanu & Fischer, 1991). This has spurred the
development of more sensitive methods for analyzing team communication.

Content analysis is a specific type of communication analysis that can be used
in conjunction with other procedures, such as protocol analysis. Its aim is to mea-
sure aspects of verbal exchange that related to team processes such as coordina-
tion. In one approach to content analysis, communication is coded with the use of
a speech act typology (Foushee & Manos, 1981). The typology can be exhaustive
or simplified to include only fundamental aspects of communication, such as the
initiating of communication and the response to communication, with each of
these two categories broken down to include types of communication (e.g., com-
mands, questions, acknowledgements; Kanki, Lozito, & Foushee, 1989). Patterns
in the types of communication can then be analyzed by calculating matrices of ini-
tiating and responding speech as well as ratios of initiating and responding speech
(Kanki et al., 1989).
Recently emerging techniques include attempts to automate the communication analysis process through the use of Latent Semantic Analysis (LSA; Keikel, Cooke, Foltz, & Shope, 2001). Another approach, Exploratory Sequential Data Analysis (Sanderson & Fisher, 1994), has been used to analyze temporal patterns in team communications (Bowers, Jentsch, Salas, & Braun, 1998). Bowers et al. (1998) coded each utterance into categories (e.g., action, response) and then analyzed the sequence in which the statements were made. This line of research linked a specific sequence of communication, closed loop communication, to higher levels of team performance. Additionally, Social Network Analysis has been used to create measures of team mental model congruity from voice (Graham, Schneider, Bauer, Bessiere, & Gonzalez, 2004) and text communications (Carley, 1997).

Event-based measurement. Event-based measurement is a general approach to designing simulation scenarios and performance measurement tools that are systematically linked to competencies targeted for training. Targeted Acceptable Responses to Generated Events (TARGETs; Fowlkes et al., 1994) is an example of event-based performance measurement in SBT. This is a structured observation methodology in which teams are exposed to scenarios consisting of contextually relevant exercises or tasks created by the researcher/trainer. These scenarios contain cues for team members to exhibit behaviors that have been identified as important for that particular team. In addition to defining the tasks considered desirable to observe, the researcher/trainer must determine what an acceptable response to the scenario is a priori by means such as subject matter expert interview, task analysis, or investigation of standard operating procedures. Acceptable responses are determined in advance so that the observer can have a checklist at the time of the observation. This significantly adds to the reliability of the observational rating. Event-based measurement, such as the TARGETs methodology, provides the opportunity to observe team behaviors that have a low frequency of occurrence in the normal operation of the team and therefore are difficult to observe in a purely naturalistic setting (Fowlkes, Dwyer, Oser, & Salas, 1998).

There have been several recent applications of the event-based approach to training (EBAT) method to SBT in health care. On the individual level, the Simulation Module for Assessment of Resident’s Targeted Event Responses (or SMARTER) tool provides a means for training emergency medicine residents (Rosen, Salas, Silvestri, Wu, & Lazzara, 2008). Additionally, this tool has been expanded to train teamwork skills as well (Rosen, Salas, Wu et al., 2008). The EBAT approach systematically links learning objectives and the competencies to be trained with the structure of the simulation scenario and performance measures. Because EBAT captures behavior during practice activities, it is particularly well suited for measuring team performance processes.

Structural knowledge assessment. Much of the theory reviewed indicates the importance of knowledge structure (e.g., expertise theory for individuals, shared mental model theory for teams). Therefore, structural knowledge assessment is a
powerful tool for understanding performance. Pathfinder is one method of accomplishing this. It is a technique related to concept mapping in the sense that both examine structural properties of knowledge representation; however, the techniques differ in terms of the methodologies used to accomplish this task (Salas, Priest et al., 2005). Essentially, Pathfinder represents knowledge in a network structure with the concept as the node and the degree of relatedness between concepts as the relations between nodes (Mohammed et al., 2000). Pathfinder then algorithmically transforms the network measures into a psychologically scaled proximity data and a representation of the underlying structure of knowledge elements (Schvaneveldt, 1990). Constructs that are rated as highly similar are visually represented with direct links; those with more tenuous relationships are represented with indirect links, and no link is present between concepts with no relationship. In addition to the visual representation of knowledge structure, Pathfinder generates a quantitative index of coherence (Salas, Priest et al., 2005). To generate these data, the researcher must first identify a list of concepts to be presented to the participants. This list is most often derived from task analysis procedures. The concepts are then presented to the participants, who make a pairwise rating of each possible pairing of concepts.

Self-report measures. Questionnaires are a popular method for assessing individual- and team-level cognitive and affective states and processes. Many measurement systems rely on the use of data collected at the individual level through self-report measures, which are then analyzed and translated into team-level characteristics (e.g., team orientation, Driskell & Salas, 1992; collective efficacy, Tagger & Seijts, 2003; team cohesion, Gully, Devine, & Whiney, 1995). These techniques involving aggregation or translation between levels require careful analysis of the constructs and the measures involved. When translating measures from one level to another, it is important to determine whether the underlying construct should be characterized as global, shared, or configural (Kozlowski & Klein, 2000). Constructs with global properties only exist on the team level (i.e., team size and structure). Constructs with shared properties are common to each member of the team (e.g., collective efficacy, trust). Constructs with configural properties are a composite of different types of individual-level properties into a team-level construct.

Automated performance recording and measurement. Many process variables, such as communication and behavior, can be automatically and continuously monitored and recorded during training sessions in a simulation. These automated measures have several advantages over traditional measurement, including being less invasive to the performance episode and containing fewer inherent biases. Automated performance recording and measurement is limited to direct and overt behavioral responses and is therefore ineffective, by itself, for measuring cognitive processes; however, types of communication analysis (e.g., sequential analysis, LSA) are being developed to address this issue (e.g., Keikel et al., 2001). The cost
of implementing automated performance recording and measurement systems can be prohibitive for many measurement purposes.

To summarize, performance measurement systems for simulation should be rooted in theoretical models of performance. Different theories will be relevant to different simulation objectives, and therefore, different measurement requirements and techniques should be applied. The simulation purpose and context should drive the selection of the theories used in the development of performance measurement systems. The previous section reviewed a sample of techniques and tools available for measurement in simulation. The following section applies these ideas to the available literature on performance measurement in simulation training and distills a set of best practices for designing and implementing measurement systems.

What Are the Best Practices in Current SBT Systems?

The nature of tasks represented by many simulations is complex. Consequently, the nature of human performance in these simulations is complex. Theories of human performance offer valuable tools for understanding performance in simulations. This understanding can then be used to generate performance measures tied to the learning objectives (the competencies targeted for training). In turn, the performance measures capture trainees’ performance during simulation and are used to generate feedback—to guide the practice activities in SBT to ensure maximum learning outcomes. However, the devil is in the details. Given the breadth of theories driving performance measurement and the variety of tools used to measure performance, the process of bringing these components together to design and implement a performance measurement system for use in a simulation training environment is complex. Therefore, in this section, we synthesize the review of theory and methods provided earlier with a review of the practice-based literature on performance measurement systems to distill a set of best practices.

Cannon-Bowers and Salas (1997) advanced a set of requirements for performance measurement in training systems. Although these requirements deal specifically with team training, they are generally applicable, with few exceptions, at the individual level as well. Therefore, we use these basic requirements of a performance measurement system as a framework for organizing the best practices distilled from the literature. These best practices are summarized in Table 3.

Performance Measurement Systems Must Consider Multiple Levels of Measurement

The multidimensional and multilevel nature of performance has been a recurring theme in the previous sections. In the simplest case, performance can be measured at the individual and team level. Adding complexity, team performance can be measured at the
team level or as an aggregate of individual-level properties (i.e., the issue of compilation vs. composition discussed earlier). Therefore, a critical factor in designing a measurement system is choosing the appropriate level of analysis for the measurement system. The type of skill being trained and the type of feedback and remediation necessary to improve performance determine the appropriateness of the level of analysis for measurement.

Table 3
Best Practices for Performance Measurement in Training Simulation

<table>
<thead>
<tr>
<th>Measurement System Requirement</th>
<th>Best Practice Performance measurement (usually) works best when . . .</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple levels of measurement</td>
<td>1. It captures <strong>multiple dimensions</strong> of performance at appropriate levels of analysis.</td>
</tr>
<tr>
<td></td>
<td>2. <strong>Event-based techniques</strong> are used to capture data at multiple levels of analysis.</td>
</tr>
<tr>
<td></td>
<td>3. <strong>Multiple measures</strong> from various sources are captured.</td>
</tr>
<tr>
<td></td>
<td>4. A systematic plan is in place to <strong>integrate data</strong> from multiple measures.</td>
</tr>
<tr>
<td>Address process as well as outcome</td>
<td>1. It captures the <strong>processes</strong> of performance.</td>
</tr>
<tr>
<td></td>
<td>2. <strong>Expert models</strong> of the task are used to compare and evaluate performance processes and outcomes.</td>
</tr>
<tr>
<td></td>
<td>3. The collection and transmission of <strong>objective</strong> measures is <strong>automated</strong>.</td>
</tr>
<tr>
<td>Describe, evaluate, and diagnose performance</td>
<td>1. Measures are <strong>descriptive</strong> of performance.</td>
</tr>
<tr>
<td></td>
<td>2. Performance can be compared to standards for <strong>desired levels of performance</strong>.</td>
</tr>
<tr>
<td></td>
<td>3. It is diagnostic-when it provides insight into the causes of performance.</td>
</tr>
<tr>
<td></td>
<td>4. It allows for <strong>performance diagnosis</strong> to be partially or fully automated.</td>
</tr>
<tr>
<td></td>
<td>5. It allows for <strong>performance evaluation</strong> to be partially or fully automated.</td>
</tr>
<tr>
<td></td>
<td>6. There is <strong>flexibility designed into embedded measures</strong> (different measures can be substituted).</td>
</tr>
<tr>
<td></td>
<td>7. It captures a <strong>broad spectrum</strong> of measures and the <strong>context of performance</strong>.</td>
</tr>
<tr>
<td></td>
<td>8. <strong>Observers are trained</strong> to high levels of reliability.</td>
</tr>
<tr>
<td></td>
<td>9. Observers use <strong>protocols</strong>.</td>
</tr>
<tr>
<td>Provide a basis for remediation</td>
<td>1. It supports learning.</td>
</tr>
<tr>
<td></td>
<td>2. It allows for the <strong>automated and manual creation of AAR aids</strong> for training remediation and feedback.</td>
</tr>
<tr>
<td></td>
<td>3. It enables <strong>automated scaffolding and performance-based coaching</strong>.</td>
</tr>
<tr>
<td></td>
<td>4. It drives real-time corrective feedback.</td>
</tr>
<tr>
<td></td>
<td>5. It is <strong>integrated with trainer controls and feedback generation</strong>.</td>
</tr>
</tbody>
</table>

Note: AAR = After Action Review.
Best Practice 1: Performance measurement usually works best when it captures **multiple dimensions** of performance at **appropriate levels** of analysis.

Performance measurement and evaluation on multiple levels is a major issue in any situation involving more than one person and especially for joint collective operations simulation (Fleming, Barto, & Johnson, 1998). Fowlkes, Dwyer, Milham, Burns, and Pierce (1999) describe the development of the Tactically Relevant Assessment of Combat Teams (TRACTs) performance measurement system for capturing performance data at multiple levels of analysis. The TRACTs system was implemented in a synthetic theater of war environment and used an event-based approach to measurement for capturing interactions constituting performance on the team and subteam levels. As discussed in the methodology section, event-based approaches such as TRACTs and TARGETs (Fowlkes et al., 1994) ease the burden on observers because observers are aware of when and where targeted individual and team behaviors would occur. The a priori specification of targeted individual- and team-level competencies of interest simplifies the process of observing multilevel multidimensional performance.

Additionally, Carolan et al. (2003) describe the development of a performance measurement system designed to capture performance in Distributed Mission Operations (DMO) training simulations. This system combines objective performance data captured by the simulation with observer ratings and information from trainees captured from posttraining questionnaires and debriefs. By combining these various methods, the measurement system provides a robust basis for evaluation and remediation, and by employing a common measurement framework, the process of integrating the various methods is facilitated. The system creates a profile of both individual and team performance with multiple methods of measurement, contributing to both levels of analysis. As one of the key purposes of performance measurement in SBT is feedback and timeliness is key to the effectiveness of feedback, there must be a plan in place for systematically integrating all of these data into a form that is meaningful and meets the goals of measurement (Oser, Cannon-Bowers, Salas, & Dwyer, 1999).

Best Practice 2: Performance measurement usually works best when **event-based techniques** are used to capture data at **multiple levels** of analysis.

Best Practice 3: Performance measurement usually works best when **multiple measures** from **various sources** are captured.

Best Practice 4: Performance measurement usually works best when a **systematic plan** is **in place** to integrate data from **multiple measures**.

Performance Measurement Systems Must Consider Processes in Addition to Outcomes

In SBT, knowledge of performance outcomes does not provide enough information to diagnose performance deficiencies, nor does it provide a basis for corrective feedback; therefore, measurement systems must capture process dimensions of
performance as well as outcomes. Outcomes alone do not directly describe the reasons that a certain level of performance is achieved. Outcome performance measures can alert trainees and trainers that a problem is present, but they provide no direction as to correcting the problem. In this sense, measurements of process are more diagnostic. Additionally, processes are more descriptive of what is to be trained.

**Best Practice 5**: Performance measurement works best when it captures the processes of performance.

Leddo, Zhang, and Pokorny (1998) report on the development of an automated performance measurement system embedded in an in-patient care simulator for medical simulation. The performance measurement system uses an expert model of the task to evaluate trainee performance behaviors during the training episode as well as trainee responses to a posttraining questionnaire presented automatically by the simulation. By combining these methods, the system is able to ensure that the trainee has developed proficiency with the task as well as a conceptual understanding of the principles being trained.

Schreiber, Watz, and Winston (2003) discuss the development of a proof of concept for a Performance Effectiveness Tracking System (PETS) for DMO training simulations. A core characteristic of the PETS is the need to capture performance process as well as outcome measures at the individual and team levels. To accomplish this, the researchers created a tool that “listens” to the data shared between various nodes in the networked simulation. The system captures multiple measures of performance by tracking the system state and trainee behavior in real time:

**Best Practice 6**: Performance measurement usually works best when expert models of the task are used as standards against which to compare and evaluate performance processes and outcomes.

**Best Practice 7**: Performance measurement usually works best when the collection and transmission of objective measures is automated.

**Performance Measurement Systems Must Describe, Evaluate, and Diagnose Performance**

By addressing the above issues (i.e., “Should the measurement system capture performance at the process or outcome phase, at the team or individual level?”), a performance measurement system is much more likely to meet its overall objectives. Generally, these objectives are the description, evaluation, and diagnosis of performance. The description of performance is conceptually simple, although in practice is often very difficult because of the complexity and dynamic nature of tasks. The evaluation of performance is the assessment of observed behaviors and judgments concerning their appropriateness. To make an evaluation of performance, the measured
performance must be compared to some criterion or standard for the desired level of performance (Oser et al., 1999). Diagnosis of performance is the determination of the causes of both effective and ineffective performance. The process of performance diagnosis involves developing a rich and detailed profile of performance that can be used to identify the presence or absence of competencies targeted for training (Salas, Rosen, Burke, Nicholson, & Howse, 2007).

Best Practice 8: Performance measurement usually works best when measures are descriptive of performance.
Best Practice 9: Performance measurement usually works best when performance can be compared to standards for desired levels of performance.
Best Practice 10: Performance measurement usually works best when it is diagnostic—when it provides insight into the causes of performance.

The expertise of the trainer is an essential component of many training simulations. However, the complexity of many training situations complicates the trainer’s task of evaluating and diagnosing performance. Oftentimes, the trainer is not in a position to measure or evaluate; their full attention is required to run the simulation. Drewes and Gonzalez (1997) present a method for automatically monitoring, evaluating, and diagnosing trainee performance during simulation training. The system—called Template Based Reasoning—uses artificial intelligence to assist in the interpretation of trainee behavior. A template consists of attributes that define the particular behavior being trained. The system monitors the trainee’s current behavior and creates a measure of confidence as to how well the current behavior matches the template. This analysis can be fed into automated feedback to the trainee as well as be used by the trainer for evaluation and remediation. A similar approach was taken by Zachary, Bilazarian, Burns, and Cannon-Bowers (1997) in the development of an embedded training system for shipboard systems. Their system used a cognitive architecture and real-time broad spectrum capture of trainee behaviors to compare what the trainee is doing to what they should be doing. This process is used to create feedback to the trainee:

Best Practice 11: Performance measurement usually works best when it allows for performance diagnosis to be partially or fully automated.
Best Practice 12: Performance measurement usually works best when it allows for performance evaluation to be partially or fully automated.

Stacy, Freeman, Lackey, and Merket (2004) discuss the development of Human Performance Objects (HPOs), a software object used for collecting performance data in federated high-level architecture simulation and training environments. HPOs enable the effective collection of performance data in distributed environments and provide a clear method for substituting various measures of performance depending
on the training needs by integrating performance measurement with the software architecture of the simulation itself. This characteristic of HPOs is extremely helpful for distributed mission training because performance measurement data are incorporated into the information shared by the various distributed simulation nodes:

*Best Practice 13:* Performance measurement usually works best when there is **flexibility designed into embedded measures** (different measures can be substituted).

M. Cooper, Viney, and McDermott (2003) describe an embedded training system within military aircraft that supports realistic in-flight training and premission rehearsal while on route to the actual mission. An important feature of this system involves mission playback. The system records all of the information available to a crew member; during a training session, a single crew member may have several different screens from which to select. After the training session, the trainer can choose the screen the trainee was using at a particular moment or any of the other screens that were available to the trainee. This ability gives the trainer access to the entire set of information available to the trainee and assists in the diagnosis of effective and ineffective performance and allows for appropriate feedback and remediation.

*Best Practice 14:* Performance measurement usually works best when it captures a **broad spectrum** of measures and the **context of performance**.

Presently, observers are still necessary components of performance measurement systems, especially when evaluating and diagnosing complex performance. When observers are used, several steps need to be taken to ensure that quality data are collected. Observers must be trained to be highly reliable in terms of internal consistency as well as interrater reliability (Baker & Salas, 1992). As noted in the methodology section, many of the tools available to observers to aid in data collection (e.g., BARS, BOS) may introduce specific biases. Training with the tools to be used during observation can reduce the method variance and yield better measurement. In addition to training, protocols need to be developed to guide the observers’ attention to salient aspects of performance and to improve the consistency and quality of data gathered during observation sessions (Fowlkes et al., 1994):

*Best Practice 15:* Performance measurement usually works best when **observers are trained** to high levels of reliability.

*Best Practice 16:* Performance measurement usually works best when **observers use protocols**.

**Performance Measurement Systems Must Provide a Basis for Remediation**

Performance diagnosis feeds into remediation, the next requirement of a performance measurement system. In a simulation training environment, the performance measurement
can be used for a number of purposes; however, the overriding purpose of measurement in simulation is to support the instructional process. Measures should inform the trainer on trainee progress, provide trainees with information of results and feedback on how to improve performance, and provide a basis for selecting future training:

*Best Practice 17*: Performance measurement usually works best when it **supports learning**.

A common form of feedback to trainees during training simulations in a military environment is the After Action Review (AAR). Traditionally, AAR systems have been add-ons to training simulations and have been criticized for resting on methodologically unsound ground (Kelly, 1999). In a brief review of the development of AAR systems in simulations, Kelly (1999) documents a trend in systems running from reporting basic outcomes (e.g., number of enemies destroyed, number of casualties, completing mission objectives) to providing more sophisticated analysis of the performance episode and tools to support changing processes involved in performance. Meliza and Brown (1996) document the Automated Training Analysis and Feedback System (ATAFS). This system was built to address the problem faced by trainers in creating AAR aids. Specifically, trainers lack the time to create effective AAR aids in a timely manner for training remediation. ATAFS was developed to ease the burden on training instructors by automating the performance measurement process during simulations as well as the construction of AAR aids for training remediation after performance episodes. The system automatically generates a bin of visual aids for AAR from performance data automatically captured by the simulation as well as facilitating manual development of visual aids by the trainer:

*Best Practice 18*: Performance measurement usually works best when it allows for the **automated and manual creation of AAR aids** for training remediation and feedback.

Additionally, Frank et al. (2004) describe the use of an automated performance measurement system embedded within a distributed training simulation. In this environment, the performance measurement system is used to automatically generate AARs for students. The AARs are used to scaffold the training—that is, adapt the training to the trainee’s performance. This has proven to be a very effective method for delivering training without the presence of trainers. Chandler and Owens (2003) report on a similar approach using intelligent tutoring systems to facilitate performance-based coaching:

*Best Practice 19*: Performance measurement usually works best when it enables **automated scaffolding** and **performance-based coaching**.

Berger (1999) discusses the use of real-time corrective feedback in the training of truck drivers. In the SIGNAL MAN system, the trainer specifies an endpoint and
path for the trainee to follow. Trainee input to the simulation is then evaluated by algorithms embedded in the training system that calculate a route-based present and desired (end) system states. A corrective feedback loop is created by displaying cues to the trainee based on the calculated route:

*Best Practice 20:* Performance measurement usually works best when it drives real-time corrective feedback.

Meliza and Paz (1996) discuss the need for training systems to integrate the feedback and control functions of a training system. In most training systems, trainers are given control over many situation or scenario variables within the simulation. By connecting the manipulations made to these variables in the simulation to the process of generating feedback (e.g., creating AARs), the workload on the trainer(s) can be greatly ameliorated. In this way, the simulation can be modified with any combination of available control variables, and the automated system measurements will provide appropriate feedback regardless of how the simulation unfolds:

*Best Practice 21:* Performance measurement usually works best when it is integrated with trainer controls and feedback generation.

**Where Do We Go From Here?**

This article has highlighted a subset of the theories of human performance used to guide measurement in SBT as well as the methodologies commonly used to capture data. Additionally, the practice-based literature has been reviewed to illustrate how theory and methods can best be applied in the development of performance measurement systems in simulation. We conclude with some comments on needs for continued research in the areas of theory, methods, and applications.

First, although theories of human performance are plentiful, there is a strong need for more precise, yet ecologically valid, theories of human performance. All too often, the robustness of implications a theory has for performance in complex environments is sacrificed for precision and vice versa. To progress, the development of performance measurement systems in SBT needs theories of human performance that can represent the complexities of interactions between trainees and the simulated environments, while simultaneously affording a detailed and precise understanding of the multiple dimensions of performance.

Second, methods need to be developed that allow for the moment-to-moment capture of data that is necessary for performance diagnosis. It is necessary to measure the processes of performance to understand the causes of performance, and therefore, a focus must be placed on the capture of data over time.
Third, applications must be developed that partially or wholly automate performance measurement, evaluation, and diagnosis. There has been advancement in the automated analysis of communication content and flow (e.g., Keikel et al., 2001), as well as the capture of task performance over time, but this is just the beginning.

Performance diagnosis requires the integration of multiple measures from multiple data sources and analyzed at multiple levels. To accomplish this in a systematic and repeatable manner requires development in the areas of theory and methods described above; however, the best practices presented in this article provide examples where designers of performance measurement systems have applied the current theories and methods in a manner that yields maximum effectiveness of performance measurement in SBT.

References


Mohammed, S., Klomski, R., & Rentsch, J. R. (2000). The measurement of team mental models: We have no shared schema. Organizational Research Methods, 3, 123-165.


**Eduardo Salas** is a Pegasus Professor of Psychology and university trustee chair at the University of Central Florida. He was editor of *Human Factors* from 2000 to 2004 and associate editor from 2004 to 2008. He is now associate editor of the *Journal of Applied Psychology*. He received a PhD in industrial/organizational psychology from Old Dominion University in 1984. **Contact:** Department of Psychology and the Institute for Simulation and Training, University of Central Florida, 3100 Technology Pkwy, Orlando, FL 32826, USA; phone: 407-882-1325; fax: 407-882-1550; e-mail: esalas@ist.ucf.edu.

**Michael A. Rosen** is a doctoral candidate in the Applied Experimental and Human Factors Psychology Program at the University of Central Florida and a graduate research associate at the Institute for Simulation and Training, where he won the student researcher of the year in 2006. He received his MA in English from the University of Central Florida in 2005. **Contact:** Department of Psychology and the Institute for Simulation and Training, University of Central Florida, 3100 Technology Pkwy, Orlando, FL 32826, USA. phone: 407-882-1408; fax: 407-882-1550; e-mail: esalas@ist.ucf.edu.

**Janet D. Held** is principal investigator for the Navy in setting aptitude standards for the Navy’s many ratings (occupations). She is the Navy’s technical representative on the joint service Military Accession Policy Working Group–Military Personnel Testing. **Contact:** Navy Personnel Research, Studies, & Technology (NPRST), Bureau of Naval Personnel (BUPERS-13), 5720 Integrity Drive Millington, TN 38055-1000, USA; phone: 901-874-4650; fax: 901-874-2570; e-mail: janet.held@navy.mil.

**Johnny J. Weissmuller** is currently the Deputy for the Headquarters Air Force Personnel Center (AFPC) Force Management Liaison Office (FMLO). He is concerned with strategic personnel analyses covering the entire military-civilian-contractor personnel spectrum and spanning from recruiting through retirement and beyond. Prior to moving to the FMLO in August 2008, he was the Chief of Test Research in the AFPC Military Testing Section with a focus on military selection, classification, and promotion testing issues. **Contact:** Deputy, Force Management Liaison Office, HQ Air Force Personnel Center, HQ AFPC/DPSF, 550 C St W, Ste 4, Randolph AFB, TX 78150-4701, USA; phone: 210-565-2238; e-mail: johnny.weissmuller@randolph.af.mil.