Chapter 2
Evaluation Criteria for Human-Automation Performance Metrics

Birsen Donmez, Patricia E. Pina, and M.L. Cummings

Abstract Previous research has identified broad metric classes for human-automation performance in order to facilitate metric selection, as well as understanding and comparing research results. However, there is still a lack of a systematic method for selecting the most efficient set of metrics when designing evaluation experiments. This chapter identifies and presents a list of evaluation criteria that can help determine the quality of a metric in terms of experimental constraints, comprehensive understanding, construct validity, statistical efficiency, and measurement technique efficiency. Based on the evaluation criteria, a comprehensive list of potential metric costs and benefits is generated. The evaluation criteria, along with the list of metric costs and benefits, and the existing generic metric classes provide a foundation for the development of a cost-benefit analysis approach that can be used for metric selection.

2.1 Introduction

Human-automation teams are common in many domains, such as command and control operations, human-robot interaction, process control, and medicine. With intelligent automation, these teams operate under a supervisory control paradigm. Supervisory control occurs when one or more human operators intermittently program and receive information from a computer that then closes an autonomous control loop through actuators and sensors of a controlled process or task environment [1]. Example applications include robotics for surgery and geologic rock sampling, and military surveillance with unmanned vehicles.

A popular metric used to evaluate human-automation performance in supervisory control is mission effectiveness [2, 3]. Mission effectiveness focuses on performance

B. Donmez (✉)
Massachusetts Institute of Technology, Department of Aeronautics and Astronautics, Cambridge, MA 02139, USA
e-mail: bdonmez@mit.edu

R. Madhavan et al. (eds.), *Performance Evaluation and Benchmarking of Intelligent Systems*, DOI 10.1007/978-1-4419-0492-8_2,

as it relates to the final output produced by the human-automation team. However, this metric fails to provide insights into the process that leads to the final mission-related output. A suboptimal process can lead to a successful completion of a mission, e.g., when humans adapt to compensate for design deficiencies. Hence, focusing on just mission effectiveness makes it difficult to extract information to detect design flaws and to design systems that can consistently support successful mission completion.

Measuring multiple human-computer system aspects such as workload and situation awareness can be valuable in diagnosing performance successes and failures, and in identifying effective training and design interventions. However, choosing an efficient set of metrics for a given experiment still remains a challenge. Many researchers select their metrics based on their past experience. Another approach to metric selection is to collect as many measures as possible to supposedly gain a comprehensive understanding of the human-automation team performance. These methods can lead to insufficient metrics, expensive experimentation and analysis, and the possibility of inflated type I errors. There appears to be a lack of a principled approach to evaluate and select the most efficient set of metrics among the large number of available metrics.

Different frameworks of metric classes are found in the literature in terms of human-autonomous vehicle interaction [4–7]. These frameworks define metric taxonomies and categorize existing metrics into high-level metric classes that assess different aspects of the human-automation team performance and are generalizable across different missions. Such frameworks can help experimenters identify system aspects that are relevant to measure. However, these frameworks do not include evaluation criteria to select specific metrics from different classes. Each metric set has advantages, limitations, and costs, thus the added value of different sets for a given context needs to be assessed to select an efficient set that maximizes value and minimizes cost.

This chapter presents a brief overview of existing generalizable metric frameworks for human-autonomous vehicle interaction and then suggests a set of evaluation criteria for metric selection. These criteria and the generic metric classes constitute the basis for the future development of a cost-benefit methodology to select supervisory control metrics.

2.2 Generalizable Metric Classes

For human-autonomous vehicle interaction, different frameworks of metric classes have been developed by researchers to facilitate metric selection, and understanding and comparison of research results. Olsen and Goodrich proposed four metric classes to measure the effectiveness of robots: task efficiency, neglect tolerance, robot attention demand, and interaction effort [4]. This set of metrics measures the individual performance of a robot, but fails to measure human performance explicitly.

Human cognitive limitations often constitute a primary bottleneck for human-automation team performance [8]. Therefore, a metric framework that can be generalized across different missions conducted by human-automation teams should include cognitive metrics to understand what drives human behavior and cognition.

In line with the idea of integrating human and automation performance metrics, Steinfeld et al. [7] suggested identifying common metrics in terms of three aspects: human, robot, and the system. Regarding human performance, the authors discussed three main metric categories: situation awareness, workload, and accuracy of mental models of device operations. This work constitutes an important effort towards developing a metric toolkit; however, this framework suffers from a lack of metrics to evaluate collaboration effectiveness among humans and among robots.

Pina et al. [5] defined a comprehensive framework for human-automation team performance based on a high-level conceptual model of human supervisory control. Figure 2.1 represents this conceptual model for a team of two humans collaborating, with each controlling an autonomous platform. The platforms also collaborate autonomously, depicted by arrows between each collaborating unit. The operators receive feedback about automation and mission performance, and adjust automation behavior through controls if required. The automation interacts with the real world through actuators and collects feedback about mission performance through sensors.

Based on this model, Pina et al. [5] defined five generalizable metric classes: mission effectiveness, automation behavior efficiency, human behavior efficiency,

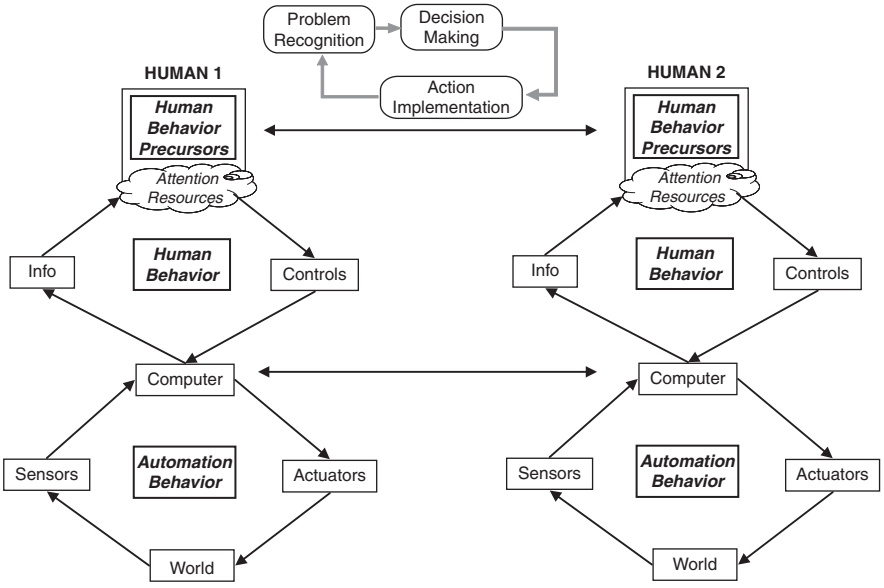

Fig. 2.1 Conceptual model of human-supervisory control (modified from Pina et al. [5])

Table 2.1 Human supervisory control metric classes [9]

Metric classes
Mission effectiveness (e.g., key mission performance parameters)
Automation behavior efficiency (e.g., usability, adequacy, autonomy, reliability)
Human Behavior Efficiency – Attention allocation efficiency (e.g., scan patterns, prioritization) – Information processing efficiency (e.g., decision making)
Human behavior precursors – Cognitive precursors (e.g., situational awareness, mental workload) – Physiological precursors (e.g., physical comfort, fatigue)
Collaborative metrics – Human/automation collaboration (e.g., trust, mental models) – Human/human collaboration (e.g., coordination efficiency, team mental model) – Automation/automation collaboration (e.g., platform's reaction time to situational events that require autonomous collaboration)

human behavior precursors, and collaborative metrics (Table 2.1). Mission effectiveness includes the previously discussed popular metrics and measures concerning how well the mission goals are achieved. Automation and human behavior efficiency measure the actions and decisions made by the individual components of the team. Human behavior precursors measure a human's internal state, including attitudes and cognitive constructs that can be the cause of and influence a given behavior. Collaborative metrics address three different aspects of team collaboration: collaboration between the human and the automation, collaboration between the humans that are in the team, and autonomous collaboration between different platforms.

These metric classes can help researchers select metrics that result in a comprehensive understanding of the human-automation performance, covering issues ranging from automation capabilities to human cognitive abilities. A rule of thumb is to select at least one metric from each metric class. However, there still is a lack of a systematic methodology to select a collection of metrics across these classes that most efficiently measures the performance of human-automation systems. The following section presents a preliminary list of evaluation criteria that can help researchers evaluate the quality of a set of metrics.

2.3 Metric Evaluation Criteria

The proposed metric evaluation criteria for human supervisory control systems consist of five general categories, listed in Table 2.2. These categories focus both on the metrics, which are constructs, and on the associated measures, which are mechanisms for expressing construct sizes. There can be multiple ways of measuring a metric. For example, situational awareness, which is a metric, can be measured based on objective or subjective measures [10]. Different measures for the same metric can generate different benefits and costs. Therefore, the criteria presented in this section evaluate a metric set by considering the metrics (e.g., situational

Table 2.2 Metric evaluation criteria

Evaluation criteria	Example
Experimental constraints	Time required to analyze a metric
Comprehensive understanding	Causal relations with other metrics
Construct validity	Power to discriminate between similar constructs
Statistical efficiency	Effect size
Measurement technique efficiency	Intrusiveness to subjects

awareness), the associated measures (e.g., subjective responses), and the measuring techniques (e.g., questionnaires given at the end of experimentation).

The costs and benefits of different research techniques in human engineering have been previously discussed in the literature [11, 12]. The list of evaluation criteria presented in this chapter is specific to the evaluation of human-automation performance and was identified through a comprehensive literature review of different metrics, measures, and measuring techniques utilized to assess human-automation interaction [9]. Advantages and disadvantages of these methods, which are discussed in detail in Pina et al. [9], fell into five general categories that constitute the proposed evaluation criteria.

These proposed criteria target human supervisory control systems, with influence from the fields of systems engineering, statistics, human factors, and psychology. These fields have their own flavors of experimental metric selection including formal design of experiment approaches such as response surface methods and factor analyses, but often which metric to select and how many are left to heuristics developed through experience.

2.3.1 Experimental Constraints

Time and monetary costs associated with measuring and analyzing a specific metric constitute the main practical considerations for metric selection. Time allocated for gathering and analyzing a metric also comes with a monetary cost due to man-hours, such as time allocated for test bed configurations. Availability of temporal and monetary resources depends on the individual project; however, resources will always be a limiting factor in all projects.

The stage of system development and the testing environment are additional factors that can guide metric selection. Early phases of system development require more controlled experimentation in order to evaluate theoretical concepts that can guide system design. Later phases of system development require a less controlled evaluation of the system in actual operation. For example, research in early phases of development can assess human behavior for different proposed automation levels, whereas research in later phases can assess the human behavior in actual operation in response to the implemented automation level.

The type of testing environment depends on available resources, safety considerations, and the stage of research development. For example, simulation

environments give researchers high experimental control, which allows them the ability to manipulate and evaluate different system design concepts accordingly. In simulation environments, researchers can create off-nominal situations and measure operator responses to such situations without exposing them to risk. However, simulation creates an artificial setting and field testing is required to assess system performance in actual use. Thus, the types of measures that can be collected are constrained by the testing environment. For example, responses to rare events are more applicable for research conducted in simulated environments, whereas observational measures can provide better value in field testing.

2.3.2 Comprehensive Understanding

It is important to maximize the understanding gained from a research study. However, due to the limited resources available, it is often not possible to collect all required metrics. Therefore, each metric should be evaluated based on how much it explains the phenomenon of interest. For example, continuous measures of workload over time (e.g., pupil dilation) can provide a more comprehensive dynamic understanding of the system compared to static, aggregate workload measures collected at the end of an experiment (e.g., subjective responses).

The most important aspect of a study is finding an answer to the primary research question. The proximity of a metric to answer the primary research question defines the importance of that metric. For example, a workload measure may not tell much without a metric to assess mission effectiveness, which is what the system designers are generally most interested in understanding. However, this does not mean that the workload measure fails to provide additional insights into the human-automation performance. Another characteristic of a metric that is important to consider is the amount of additional understanding gained using a specific metric when a set of metrics are collected. For example, rather than having two metrics from one metric class (e.g., mission effectiveness), having one metric from two different metric classes (e.g., mission effectiveness and human behavior) can provide a better understanding of human-automation performance.

In addition to providing additional understanding, another desired metric quality is its causal relations with other metrics. A better understanding can be gained if a metric can help explain other metrics' outcomes. For example, operator response to an event, hence human behavior, will often be dependent on the conditions and/or the operator's internal state when the event occurs. The response to an event can be described in terms of three set of variables [13]: a pre-event phase that defines how the operator adapts to the environment; an event-response phase that describes the operator's behavior in accommodating the event; and an outcome phase that describes the outcome of the response process. The underlying reasons for the operator's behavior and the final outcome of an event can be better understood if the initial conditions and operator's state when the event occurs are also measured. When used as covariates in statistical analysis, the initial conditions of the environment and the operator can help explain the variability in other metrics of interest. Thus,

in addition to human behavior, experimenters are encouraged to measure human behavior precursors in order to assess the operator state and environmental conditions, which may influence human behavior.

High correlation between different measures, even if they are intended to assess different metrics, is another limiting factor in metric/measure selection. A high correlation can be indicative of the fact that multiple measures can assess the same metric or the same phenomenon. Hence, including multiple measures that are highly correlated with each other can result in wasted resources and also bring into question construct validity, which is discussed next.

2.3.3 Construct Validity

Construct validity refers to how well the associated measure captures the metric or construct of interest. For example, subjective measures of situational awareness ask subjects to rate the amount of situational awareness they had on a given scenario or task. These measures are proposed to help in understanding subjects' situational awareness [10, 14]. However, self-ratings assess meta-comprehension rather than comprehension of the situation: it is unclear whether operators are aware of their lack of situational awareness. Therefore, subjective responses on situational awareness are not valid to assess actual situational awareness, but rather the awareness of lack of situational awareness.

Good construct validity requires a measure to have high sensitivity to changes in the targeted construct. That is, the measure should reflect the change as the construct moves from low to high levels [15]. For example, primary task performance generally starts to break down when the workload reaches higher levels [15, 16]. Therefore, primary task performance measures are not sensitive to changes in the workload at lower workload levels, since with sufficient spare processing capacity, operators are able to compensate for the increase in workload.

A measure with high construct validity should also be able to discriminate between similar constructs. The power to discriminate between similar constructs is especially important for abstract constructs that are hard to measure and difficult to define, such as situational awareness or attentiveness. An example measure that fails to discriminate two related metrics is galvanic skin response. Galvanic skin response is the change in electrical conductance of the skin attributable to the stimulation of the sympathetic nervous system and the production of sweat. Perspiration causes an increase in skin conductance, thus galvanic skin response has been proposed and used to measure workload and stress levels (e.g., [17]). However, even if workload and stress are related, they still are two separate metrics. Therefore, galvanic skin response alone cannot suggest a change in workload.

Good construct validity also requires the selected measure to have high inter- and intra-subject reliability. Inter-subject reliability requires the measure to assess the same construct for every subject, whereas intra-subject reliability requires the measure to assess the same construct if the measure was repeatedly collected from the same subject under identical conditions.

Intra- and inter-subject reliabilities are especially of concern for subjective measures. For example, self-ratings are widely utilized for mental workload assessment [18, 19]. This technique requires operators to rate the workload or effort experienced while performing a task or a mission. Self-ratings are easy to administer, non-intrusive, and inexpensive. However, different individuals may have different interpretations of workload, leading to decreased inter-subject reliability. For example, some participants may not be able to separate mental workload from physical workload [20], and some participants may report their peak workload, whereas others may report their average workload. Another example of low inter-subject reliability is for subjective measures of situational awareness. Vidulich and Hughes [10] found that about half of their participants rated situational awareness by gauging the amount of information to which they attended; while the other half of the participants rated their SA by gauging the amount of information they thought they had overlooked. Participants may also have recall problems if the subjective ratings are collected at the end of a test period, raising concerns on the intra-subject reliability of subjective measures.

2.3.4 Statistical Efficiency

There are three metric qualities that should be considered to ensure statistical efficiency: total number of measures collected, frequency of observations, and effect size.

Analyzing multiple measures inflates type I error. That is, as more dependent variables are analyzed, finding a significant effect when there is none becomes more likely. The inflation of type I error due to multiple dependent variables can be handled with multivariate analysis techniques, such as Multivariate Analysis of Variance (MANOVA) [21]. However, it should be noted that multivariate analyses are harder to conduct, as researchers are more prone to include irrelevant variables in multivariate analyses, possibly hiding the few significant differences among many insignificant ones. The best way to avoid failure to identify significant differences is to design an effective experiment with the most parsimonious metric/measure set that specifically addresses the research question.

Another metric characteristic that needs to be considered is the frequency of observations required for statistical analysis. Supervisory control applications require humans to be monitors of automated systems, with intermittent interaction. Because humans are poor monitors by nature [22], human monitoring efficiency is an important metric to measure in many applications. The problem with assessing monitoring efficiency is that, in most domains, errors or critical signals are rare, and operators can have an entire career without encountering them. For that reason, in order to have a realistic experiment, such rare events cannot be included in a study with sufficient frequency. Therefore, if a metric requires response to rare events, the associated number of observations may not enable the researchers to extract meaningful information from this metric. Moreover, observed events with a low frequency of occurrence cannot be statistically analyzed unless data is obtained

from a very large number of subjects, such as in medical studies on rare diseases. Conducting such large scale supervisory control experiments is generally cost-prohibitive.

The number of subjects that can be recruited for a study is especially limited when participants are domain experts such as pilots. The power to identify a significant difference, when there is one, depends on the differences in the means of factor levels and the standard errors of these means, which constitute the effect size. Standard errors of the means are determined by the number of subjects. One way to compensate for limited number of subjects in a study is to use more sensitive measures that will provide a large separation between different conditions, that is, a high effect size. Experimental power can also be increased by reducing error variance by collecting repeated measures on subjects, focusing on sub-populations (e.g., experienced pilots), and/or increasing the magnitude of manipulation for independent variables (low and high intensity rather than low and medium intensity). However, it should also be noted that increased experimental control, such as using sub-populations, can lead to less generalizable results, and there is a tradeoff between the two.

2.3.5 Measurement Technique Efficiency

The data collection technique associated with a specific metric should not be intrusive to the subjects or to the nature of the task. For example, eye trackers are used for capturing operators' visual attention [23, 24]. However, head-mounted eye trackers can be uncomfortable for the subjects, and hence influence their responses. Wearing an eye-tracker can also lead to an unrealistic situation that is not representative of the task performed in the real world.

Eye trackers are an example of how a measurement instrument can interfere with the nature of the task. The measuring technique itself can also interfere with the realism of the study. For example, off-line query methods are used to measure operators' situational awareness [25]. These methods are based on briefly halting the experiment at randomly selected intervals, blanking the displays, and administering a battery of queries to the operators. This situational awareness measure assesses global situational awareness by calculating the accuracy of an operator's responses. The collection of the measure requires the interruption of the task in a way that is unrepresentative of real operating conditions. The interruption may also interfere with other metrics such as operator's performance and workload, as well as other temporal-based metrics.

2.4 Metric Costs vs. Benefits

The evaluation criteria discussed previously can be translated into potential cost-benefit parameters as seen in Table 2.3, which can be ultimately used to define cost and benefit functions of a metric set for a given experiment. The breakdown in

Table 2.3 Representative cost-benefit parameters for metric selection

Costs		
Data gathering	Preparation	Time to setup Expertise required
	Data collection	Equipment Time Measurement error likelihood
	Subject recruitment	Compensation IRB preparation and submission Time spent recruiting subjects
Data analysis	Data storage/transfer	Equipment Time
	Data reduction	Time Expertise required Software
	Statistical analysis	Error proneness given the required expertise Time Software Expertise
Benefits		
Comprehensive understanding		Proximity to primary research question Coverage – Additional understanding given other metrics Causal relations to other metrics
Construct validity		Sensitivity Power to discriminate between similar constructs Inter-subject reliability Intra-subject reliability
Statistical efficiency		Effect size Difference in means / Error variance Frequency of observations Total number of measures collected
Measurement technique efficiency		Non-intrusiveness to subjects Non-intrusiveness to task nature
Appropriateness for system development phase/testing environment		

Table 2.3 is based on the ability to assign a monetary cost to an item. Parameters listed as cost items can be assigned a monetary cost, whereas the parameters listed as benefit items cannot be assigned a monetary cost but nonetheless can be expressed in some kind of a utility function. However, some of the parameters listed under benefits can also be considered as potential costs in non-monetary terms, leading to a negative benefit.

It should be noted that the entries in Table 2.3 are not independent of each other, and tradeoffs exist. For example, recruiting experienced subjects can enhance con-

struct validity and statistical efficiency, however, this may be more time consuming. Figure 2.2 presents results of an experiment conducted to evaluate an automated navigation path planning algorithm in comparison to manual path planning using paper charts in terms of time to generate a plan [26]. Two groups of subjects were recruited for this experiment: civilian and military. The variability of responses of the military group was less than the civilian group, resulting in smaller error variance and larger effect size. However, recruiting military participants requires more effort as these participants are more specialized. Such tradeoffs need to be evaluated by individual researchers based on their specific research objectives and available resources.

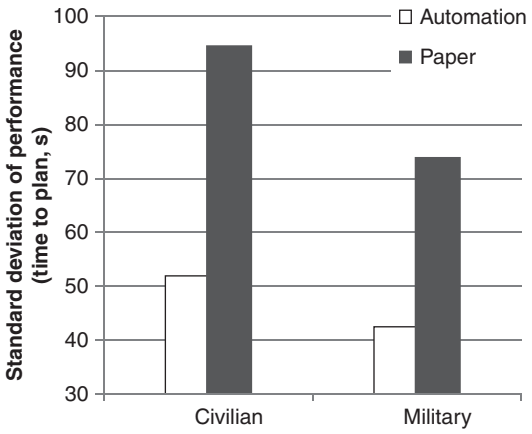

Fig. 2.2 Data variability for different subject populations

In order to demonstrate how metrics, measures, and measurement techniques can be evaluated using Table 2.3 as a guideline, the following sections present two human behavior metrics, i.e., mental workload and attention allocation efficiency, as examples for evaluating different measures.

2.4.1 Example 1: Mental Workload Measures

Workload is a result of the demands a task imposes on the operator's limited resources. Thus, workload is not only task-specific, but also person-specific. The measurement of mental workload enables, for example, identification of bottlenecks in the system or the mission in which performance can be negatively impacted. Mental workload measures can be classified into three main categories: performance, subjective, and physiological (Table 2.4). This section presents the limitations and advantages associated with each measure guided by Table 2.3. The discussions are summarized in Table 2.5.

Table 2.4 Example measures of mental workload

Measures		Techniques
Performance	Speed or accuracy for the primary task	Primary task
	Time to respond to messages through an embedded chat interface	Secondary task
Subjective (self-ratings)	Modified Cooper-Harper Scale for workload	Unidimensional questionnaires
	NASA TLX	Multidimensional questionnaires
Physiological	Blink frequency	Eye tracking
	Pupil diameter	Eye tracking
	Heart rate variability coefficient	Electrocardiogram
	Amplitudes of the N100 and P300 components of the event-related potential	Electroencephalogram
	Skin electrical conductance	Galvanic skin response

2.4.1.1 Performance Measures

Performance measures are based on the principle that workload is inversely related to the level of task performance [27]. Primary task performance should always be studied in any experiment, thus, utilizing it to assess workload comes with no additional cost or effort. However, this measure presents severe limitations as a mental workload metric, especially in terms of construct validity. Primary task performance is only sensitive in the "overload" region, when the task demands more resources from the operator than are available. Thus, it does not discriminate between two primary tasks in the "underload" region (i.e., the operator has sufficient reserve capacity to reach perfect performance). In addition, primary task performance is not only affected by workload levels, but also by other factors such as correctness of the decisions made by the operator.

Secondary task performance as a workload measure can help researchers assess the amount of residual attention an operator would have in case of an unexpected system failure or event requiring operator intervention [28]. Therefore, it provides additional coverage for understanding human-automation performance. Secondary task measures are also sensitive to differences in primary task demands that may not be reflected in primary task performance, so have better construct validity. However, in order to achieve good construct validity, a secondary task should be selected with specific attention to the types of resources it requires. Humans have different types of resources (e.g., perceptual resources for visual signals vs. perceptual resources for auditory signals) [20]. Therefore, workload resulting from the primary task can be greatly underestimated if the resource demands of the secondary task do not match those of the primary task.

Table 2.5 Evaluation of workload measures

Measures	Advantages	Limitations
Primary task performance	*Cost:* – Can require major cost/effort. However, no additional cost/effort required if already collected to assess mission effectiveness. *Comprehensive Understanding:* – High proximity to primary research question	*Construct Validity:* – Insensitive in the "underload" region – Affected by other factors
Secondary task performance	*Comprehensive Understanding:* – Coverage (assesses the residual attention an operator has) *Construct Validity:* – Sensitivity	*Cost:* – Some level of additional cost/effort *Measurement Technique Efficiency:* – Intrusive to task nature (if not representative of the real task)
Subjective measures	*Cost:* – Cheap equipment, easy to administer *Measurement Technique Efficiency:* – Not intrusive to subjects or the task	*Cost:* – More expertise required for data analysis – More subjects required to achieve adequate power *Construct Validity:* – Inter-subject reliability – Intra-subject reliability – Power to discriminate between similar constructs *Statistical Efficiency:* – Large number of observations required
Physiological measures	*Comprehensive Understanding:* – Continuous, real-time measure	*Cost:* – High level of equipment cost and expertise required – Data analysis is time consuming and requires expertise – Measurement error likelihood *Construct Validity:* – Power to discriminate between similar constructs *Measurement Technique Efficiency:* – Intrusive to subjects and task nature *Appropriateness for system development phase:* – Typically appropriate only for laboratory settings

Some of the secondary tasks that have been proposed and employed include producing finger or foot taps at a constant rate, generating random numbers, or reacting to a secondary-task stimulus [27]. Secondary tasks that are not representative of operator's real tasks may interfere with and disrupt performance of the primary task. However, problems with intrusiveness can be mitigated if embedded secondary tasks are used. In those cases, the secondary task is part of operators' responsibilities but has lower priority in the task hierarchy than the primary task. For example, Cummings and Guerlain used a chat interface as an embedded secondary task measurement tool [29]. Creating an embedded secondary task resolves the issues related to intrusiveness, however, it also requires a larger developmental cost and effort.

2.4.1.2 Subjective Measures

Subjective measures require operators to rate the workload or effort experienced while performing a task or a mission. Unidimensional scale techniques involve asking the participant for a rating of overall workload for a given task condition or at a given point in time [18, 30]. Multidimensional scale techniques require the operator to rate various characteristics of perceived workload [19, 31], and generally possess better diagnostic abilities than the unidimensional scale techniques. Self-ratings have been widely utilized for workload assessment, most likely due to their ease of use. Additional advantages are their non-intrusive nature and low cost. Disadvantages include recall problems, and the variability of workload interpretations between different individuals. In addition, it is unclear whether subjects' reported workload correlates with peak or average workload level. Another potential problem is the difficulty that humans can have when introspectively diagnosing a multidimensional construct, and in particular, separating workload elements [20]. Moreover, self-ratings measure perceived workload rather than actual workload. However, understanding how workload is perceived can be sometimes as important as measuring actual workload.

Self-ratings are generally assessed using a Likert scale that generates ordinal data. The statistical analysis appropriate for such data (e.g., logistic regression, non-parametric methods) requires more expertise than simply conducting analysis of variance (ANOVA). Moreover, the number of subjects needed to reach adequate statistical power for this type of analysis is much higher than it is for ANOVA. Thus, even if subjective measures are low cost during the experimental preparation phase, they may impose substantial costs later by requiring additional expertise for data analysis as well as additional data collection.

2.4.1.3 Physiological Measures

Physiological measures such as heart rate variability, eye movement activity, and galvanic skin response are indicative of operators' level of effort and engagement, and have also been used to assess operator workload. Findings indicate that blink rate, blink duration, and saccade duration all decrease with increased workload, while pupil diameter, number of saccades, and the frequency of long fixations all increase [32]. Heart rate variability is generally found to decrease as workload

increases [33]. The electroencephalogram (EEG) has been shown to reflect subtle shifts in workload. However, it also reflects subtle shifts in alertness and attention, which are related to workload, but can reflect different effects. In addition, significant correlations between EEG indices of cognitive state changes and performance have been reported [34–36]. As discussed previously, galvanic skin response (GSR) can be indicative of workload, as well as stress levels [17].

It is important to note that none of these physiological measures directly assess workload. These measures are sensitive to changes in stress, alertness, or attention, and it is almost impossible to discriminate whether the physiological parameters vary as a consequence of mental workload or due to other factors. Thus, the construct validity of physiological measures to assess workload is questionable.

An advantage of physiological measures is the potential for a continuous, real-time measure of ongoing operator states. Such a comprehensive understanding of operator workload can enable researchers to optimize operator workload, using times of inactivity to schedule less critical tasks or deliver non-critical messages so that they do not accumulate during peak periods [37]. Moreover, this type of knowledge could be used to adapt automation, with automation taking on more responsibilities during high operator workload [38].

Some additional problems associated with physiological measures are sensor noise (i.e., high levels of measurement error likelihood), high equipment cost, intrusiveness to task nature and subjects, and the level of expertise as well as additional time required to setup the experiment, collect data, and analyze data. Moreover, due to the significant effort that goes into setting up and calibrating the equipment, physiological measures are very difficult to use outside of laboratory settings.

2.4.2 Example 2: Attention Allocation Efficiency Measures

In supervisory control applications, operators supervise and divide their attentiveness across a series of dynamic processes, sampling information from different channels and looking for critical events. Evaluating attention allocation efficiency involves not only assessing if operators know where to find the information or the functionality they need, but also if they know when to look for a given piece of information or when to execute a given function [39]. Attention allocation measures aid in the understanding of whether and how a particular element on the display is effectively used by the operators. In addition, attention allocation efficiency measures also assess operators' strategies and priorities. It should be noted that some researchers are interested in comparing actual attention allocation strategies with optimal strategies; however, optimal strategies might ultimately be impossible to know. In some cases, it might be possible to approximate optimal strategies via dynamic programming or some other optimization technique [40]. Otherwise, the expert operators' strategy or the best performer's strategy can be used for comparison.

As shown in Table 2.6, there are three main approaches to study attention allocation: eye movements, hand movements, and verbal protocols. Table 2.7 presents the limitations and advantages associated with different measures in terms of the cost-benefit parameters identified in Table 2.3.

Table 2.6 Example attention allocation efficiency measures

Measures	Techniques
Proportion of time that the visual gaze is within each "area of interest" of an interface	Eye tracking
Average number of visits per min to each "area of interest" of an interface	Human interface-inputs
Switching time for multiple tasks	Human interface-inputs
Information used	Human interface-inputs
Operators' task and event priority hierarchies	Verbal protocols

Extensive research has been conducted with eye trackers and video cameras to infer operators' attention allocation strategies based on the assumption that the length and the frequency of eye fixations on a specific display element indicate the level of attention on the element [39, 41]. Attention allocation metrics based on eye movement activity can be dwell time (or glance duration) and glance frequency spent within each "area of interest" of the interface. While visual resources are not the only human resources available, as information acquisition typically occurs through vision in supervisory control settings, visual attention can be used to infer operators' strategies and the employment of cognitive resources. Eye tracking to assess attention allocation efficiency comes with similar limitations to physiological measures used for workload assessment, which have been discussed in Section 2.4.1.

The human interface-inputs reflect operators' physical actions, which are the result of the operators' cognitive processes. Thus operators' mouse clicking can be used to measure operators' actions, determine what information was used, and to infer operators' cognitive strategies [23, 42]. A general limitation with capturing human interface-inputs is that directing attention does not necessarily result in an immediate action, so inferring attention allocation in this manner could be subject to missing states.

Verbal protocols require operators to verbally describe their thoughts, strategies, and decisions, and can be employed simultaneously while operators perform a task, or retrospectively after a task is completed. Verbal protocols are usually videotaped so that researchers can compare what subjects say, while simultaneously observing the system state through the interface the subjects used. This technique provides insights into operators' priorities and decision making strategies, but it can be time consuming and is highly dependent on operators' verbal skills and memory. Moreover, if the operator is interrupted while performing a task, verbal protocols can be intrusive to the task.

2.5 Discussion

Supervisory control of automation is a complex phenomenon with high levels of uncertainty, time-pressure, and a dynamic environment. The performance of human-automation teams depends on multiple components such as human behavior,

Table 2.7 Evaluation of different attention allocation efficiency measures

Measures	Advantages	Limitations
Eye movements (eye tracking)	*Comprehensive Understanding:* – Continuous measure of visual attention allocation	*Cost:* – High level of equipment cost and expertise required – Data analysis is time consuming and requires expertise – Measurement error likelihood *Construct Validity:* – Limited correlation between gaze and thinking *Measurement Technique Efficiency:* – Intrusive to subjects and task nature *Appropriateness for System Development Phase:* – Appropriate for laboratory settings
Interface clicks (human interface-inputs)	*Comprehensive Understanding:* – Continuous measure of subjects' actions	*Cost:* – Time consuming during data analysis *Construct Validity:* – Directing attention does not always result in an immediate interface action
Subjective measures (verbal protocols)	*Comprehensive Understanding:* – Insight into operators' priorities and decision making strategies	*Cost:* – Time intensive *Construct Validity:* – Inter-subject reliability (dependent on operator's verbal skills) – Intra-subject reliability (recall problems with retrospective protocols) *Measurement Technique Efficiency:* – Intrusive to task nature (interference problems with real-time protocols) *Appropriateness for System Development Phase:* – Appropriate for laboratory settings

automation behavior, human cognitive and physical capabilities, team interactions, etc. Because of the complex nature of supervisory control, there are many different metrics that can be utilized to assess performance. However, it is not feasible to collect all possible metrics. Moreover, collecting multiple metrics that are correlated can lead to statistical problems such as inflated type I errors.

This chapter presented a list of evaluation criteria and cost-benefit parameters based on the criteria for determining a set of metrics for a given supervisory control research question. Thus, a limitation of this list of evaluation criteria is that it is not comprehensive enough to address all issues relevant to assessing human-technology interactions. The most prominent issues for assessing human-automation interaction were identified through a comprehensive literature review [9] and were populated under five major categories: experimental constraints, comprehensive understanding, construct validity, statistical efficiency, and measurement technique efficiency. It should be noted that there are interactions between these major categories. For example, the intrusiveness of a given measuring technique can affect the construct validity for a different metric. In one such case, if situational awareness is measured by halting the experiment and querying the operator, then the construct validity for the mission effectiveness or human behavior metrics become questionable. Therefore, the evaluation criteria presented in this chapter should be applied to a collection of metrics rather than each individual metric, taking the interactions between different metrics into consideration.

The list of evaluation criteria and the relevant cost-benefit parameters presented in this chapter are guidelines for metric selection. It should be noted that there is not a single set of metrics that are the most efficient across all applications. The specific research aspects such as available resources and the questions of interest will ultimately determine the relative metric quality. Moreover, depending on the specific research objectives and limitations, the cost-benefit parameters presented in Table 2.3 can have different levels of importance. Thus, these parameters can receive a range of weights in cost-benefit functions created for different applications. Identifying the most appropriate technique for helping researchers to assign their subjective weights is under investigation as part of an ongoing research effort. Thus, future research will further develop this cost-benefit analysis approach, which will systematically identify an efficient set of metrics for classifications of research studies.

Acknowledgments This research was funded by the US Army Aberdeen Test Center. The authors would like to thank Dr. Heecheon You for reviewing the manuscript.

References

1. T. B. Sheridan, *Telerobotics, Automation, and Human Supervisory Control*. Cambridge, MA: The MIT Press, 1992.
2. J. Scholtz, J. Young, J. L. Drury, and H. A. Yanco, “Evaluation of human-robot interaction awareness in search and rescue,” in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. New Orleans, 2004.
3. N. J. Cooke, E. Salas, P. A. Kiekel, and B. Bell, “Advances in measuring team cognition,” in *Team Cognition: Understanding the Factors that Drive Process and Performance*, E. Salas and S. M. Fiore, Eds. Washington, D.C.: American Psychological Association, 2004, pp. 83–106.
4. R. O. Olsen and M. A. Goodrich, “Metrics for evaluating human-robot interactions,” in *Proceedings of NIST Performance Metrics for Intelligent Systems Workshop*, 2003.

5. P. E. Pina, M. L. Cummings, J. W. Crandall, and M. Della Penna, "Identifying generalizable metric classes to evaluate human-robot teams," in *Proceedings of Metrics for Human-Robot Interaction Workshop at the 3rd Annual Conference on Human-Robot Interaction.* Amsterdam, The Netherlands, 2008.
6. J. W. Crandall and M. L. Cummings, "Identifying predictive metrics for supervisory control of multiple robots," *IEEE Transactions on Robotics – Special Issue on Human-Robot Interaction*, vol. 23, pp. 942-951, 2007.
7. A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. A. Goodrich, "Common metrics for human-robot interaction," in *Proceedings of the 1st Annual IEEE/ACM Conference on Human Robot Interaction (Salt Lake City, Utah).* New York, NY: ACM Press, 2006.
8. C. D. Wickens, J. D. Lee, Y. Liu, and S. G. Becker, *An Introduction to Human Factors Engineering*, 2nd ed. Upper Saddle River, New Jersey: Pearson Education, Inc., 2004.
9. P. E. Pina, B. Donmez, and M. L. Cummings, *Selecting Metrics to Evaluate Human Supervisory Control Applications*, MIT Humans and Automation Laboratory, Cambridge, MA HAL2008-04, 2008.
10. M. A. Vidulich and E. R. Hughes, "Testing a subjective metric of situation awareness," in *Proceedings of the Human Factors Society 35th Annual Meeting*. Santa Monica, CA: The Human Factors and Ergonomics Society, 1991, pp. 1307–1311.
11. A. Chapanis, *Research Techniques in Human Engineering*. Baltimore: The Johns Hopkins Press, 1965.
12. M. S. Sanders and E. J. McCormick, *Human Factors in Engineering and Design*. New York: McGraw-Hill, 1993.
13. B. Donmez, L. Boyle, and J. D. Lee, "The impact of distraction mitigation strategies on driving performance," *Human Factors*, vol. 48, pp. 785–804, 2006.
14. R. M. Taylor, "Situational awareness rating technique (SART): the development of a tool for aircrew systems design," in *Proceedings of the NATO Advisory Group for Aerospace Research and Development (AGARD) Situational Awareness in Aerospace Operations Symposium (AGARD-CP-478)*, 1989, p. 17.
15. F. T. Eggemeier, C. A. Shingledecker, and M. S. Crabtree, "Workload measurement in system design and evalution," in *Proceeding of the Human Factors Society 29th Annual Meeting*. Baltimore, MD, 1985, pp. 215–219.
16. F. T. Eggemeier, M. S. Crabtree, and P. A. LaPoint, "The effect of delayed report on subjective ratings of mental workload," in *Proceedings of the Human Factors Society 27th Annual Meeting*. Norfolk, VA, 1983, pp. 139–143.
17. S. Levin, D. J. France, R. Hemphill, I. Jones, K. Y. Chen, D. Ricard, R. Makowski, and D. Aronsky, "Tracking workload in the emergency department," *Human Factors*, vol. 48, pp. 526–539, 2006.
18. W. W. Wierwille and J. G. Casali, "A validated rating scale for global mental workload measurement applications," in *Proceedings of the Human Factors Society 27th Annual Meeting*. Santa Monica, CA, 1983, pp. 129–133.
19. S. G. Hart and L. E. Staveland, "Development of NASA-TLX (Task Load Index): results of empirical and theoretical research," in *Human Mental Workload*, P. Hancock and N. Meshkati, Eds. Amsterdam, The Netherlands: North Holland B.V., 1988, pp. 139–183.
20. R. D. O'Donnell and F. T. Eggemeier, "Workload assessment methodology," in *Handbook of Perception and Human Performance: Vol. II. Cognitive Processes and Performance*, K. R. Boff, L. Kaufmann, and J. P. Thomas, Eds. New York: Wiley Interscience, 1986, pp. 42-1–42-49.
21. R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, 5th ed. NJ: Pearson Education, 2002.
22. T. B. Sheridan, *Humans and Automation: System Design and Research Issues*. New York, NY: John Wiley & Sons Inc., 2002.
23. M. E. Janzen and K. J. Vicente, "Attention allocation within the abstraction hierarchy," *International Journal of Human-Computer Studies*, vol. 48, pp. 521–545, 1998.

24. B. Donmez, L. Boyle, and J. D. Lee, "Safety implications of providing real-time feedback to distracted drivers," *Accident Analysis & Prevention*, vol. 39, pp. 581–590, 2007.
25. M. R. Endsley, B. Bolte, and D. G. Jones, *Designing for Situation Awareness: An Approach to User-Centered Design*. Boca Raton, FL: CRC Press, Taylor & Francis Group, 2003.
26. M. Buchin, "Assessing the impact of automated path planning aids in the maritime community," in *Electrical Engineering and Computer Science M.Eng.* Cambridge, MA: Massachusetts Institute of Technology, 2009.
27. C. D. Wickens and J. G. Hollands, *Engineering Psychology and Human Performance*, 3rd ed. New Jersey: Prentice Hall, 1999.
28. G. D. Ogden, J. M. Levine, and E. J. Eisner, "Measurement of workload by secondary tasks," *Human Factors*, vol. 21, pp. 529–548, 1979.
29. M. L. Cummings and S. Guerlain, "Using a chat interface as an embedded secondary tasking tool," in *Proceedings of the 2nd Annual Human Performance, Situation Awareness, and Automation Technology Conference*. Daytona Beach, FL, 2004.
30. A. H. Roscoe and G. A. Ellis, *A Subjective Rating Scale for Assessing Pilot Workload in Flight: A Decade of Practical Use*, Royal Aeronautical Establishment, Farnborough, England TR90019, 1990.
31. G. B. Reid and T. E. Nygren, "The subjective workload assessment technique: a scaling procedure for measuring mental workload," in *Human Mental Workload*, P. Hancock and N. Meshkati, Eds. Amsterdam, The Netherlands: North Holland, 1988, pp. 185–218.
32. U. Ahlstrom and F. Friedman-Berg, *Subjective Workload Ratings and Eye Movement Activity Measures*, US Department of Transportation, Federal Aviation Administration DOT/FAA/ACT-05/32, 2005.
33. A. J. Tattersall and G. R. J. Hockey, "Level of operator control and changes in heart rate variability during simulated flight maintenance," *Human Factors*, vol. 37, pp. 682–698, 1995.
34. C. Berka, D. J. Levendowski, M. Cventovic, M. M. Petrovic, G. F. Davis, M. N. Lumicao, M. V. Popovic, V. T. Zivkovic, R. E. Olmstead, and P. Westbrook, "Real-time analysis of EEG indices of alertness, cognition, and memory acquired with a wireless EEG headset," *International Journal of Human Computer Interaction*, vol. 17, pp. 151–170, 2004.
35. J. B. Brookings, G. F. Wilson, and C. R. Swain, "Psychophysiological responses to changes in workload during simulated air-traffic control," *Biological Psychology*, vol. 42, pp. 361–377, 1996.
36. K. A. Brookhuis and D. De Waard, "The use of psychophysiology to assess driver status," *Ergonomics*, vol. 36, 1993.
37. S. T. Iqbal, P. D. Adamczyk, S. Zheng, and B. P. Bailey, "Towards an index of opportunity: understanding changes in mental workload during task execution," in *Proceedings of the ACM Conference on Human Factors in Computing Systems.* Portland, Oregon, 2005, pp. 311–320.
38. R. Parasuraman and P. A. Hancock, "Adaptive control of mental workload," in *Stress, Workload, and Fatigue*, P. A. Hancock and P. A. Desmond, Eds. Mahwah, New Jersey: Lawrence Erlbaum Associates, Publishers, 2001, pp. 305–320.
39. D. A. Talluer and C. D. Wickens, "The effect of pilot visual scanning strategies on traffic detection accuracy and aircraft control," in *Proceedings of the 12th International Symposium on Aviation Psychology.* Dayton, OH, 2003.
40. M. Puterman, *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. New Jersey: Wiley, 2005.
41. C. D. Wickens, J. Helleberg, J. Goh, Xu, X., and W. J. Horrey, *Pilot Task Management: Testing and Attentional Expected Value Model of Visual Scanning*, NASA Ames Research Center, Moffett Field, CA ARL-01-14/NASA-01-7, 2001.
42. S. Bruni, J. Marquez, A. S. Brzezinski, and M. L. Cummings, "Visualizing operators' cognitive strategies in multivariate optimization," in *Proceedings of the Human Factors and Ergonomics Society's 50th Annual Meeting.* San Francisco, CA, 2006.

http://www.springer.com/978-1-4419-0491-1

Performance Evaluation and Benchmarking of Intelligent Systems

Madhavan, R.; Tunstel, E.; Messina, E. (Eds.)
2009, XIX, 338 p., Hardcover
ISBN: 978-1-4419-0491-1