Evaluation Criteria for Human-Automation Performance Metrics

Birsen Donmez
MIT
Dept. of Aero-Astro
Cambridge, MA, USA
1(617)258-5046
bdonmez@mit.edu

Patricia E. Pina
MIT
Dept. of Aero-Astro
Cambridge, MA, USA
1(617)258-5046
ppina@mit.edu

M. L. Cummings
MIT
Dept. of Aero-Astro
Cambridge, MA, USA
1(617)252-1512
missyc@mit.edu

ABSTRACT

Previous research has identified broad metric classes for human-automation performance to facilitate metric selection, as well as understanding and comparison of research results. However, there is still lack of an objective method for selecting the most efficient set of metrics. This research 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. Future research will build on these evaluation criteria and existing generic metric classes to develop a cost-benefit analysis approach that can be used for metric selection.

Categories and Subject Descriptors

C.4 [**Performance of Systems**]: Measurement Techniques
J.7 [**Computers in Other Systems**]

General Terms

Measurement, Performance, Experimentation, Human Factors, Standardization, Theory.

Keywords

Metric Quality, Human Supervisory Control, Validity, Statistics, Experiments.

1. INTRODUCTION

Human-automation teams are common in many domains, such as military operations, process control, and medicine. With intelligent automation, these teams operate under a supervisory control paradigm. “Supervisory control means that one or more human operators are intermittently programming and continually receiving information from a computer that itself closes an autonomous control loop through artificial effectors and sensors to the controlled process or task environment [1].” Example applications include robotics for surgery, rock sampling for geology research, 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 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 the 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 the situational awareness of the human, can be valuable in diagnosing performance successes and failures, and 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 the set that maximizes value and minimizes cost.

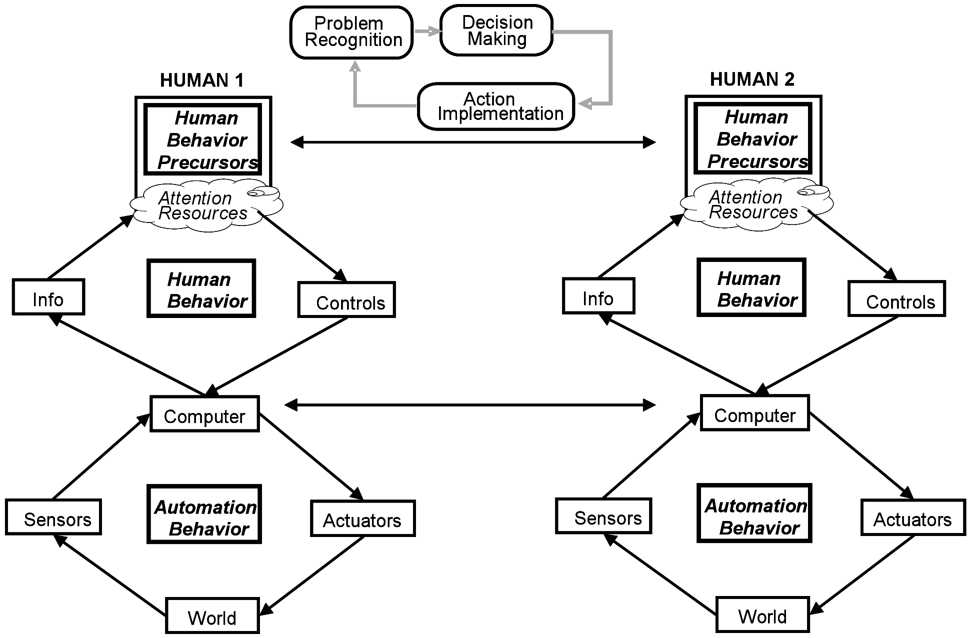

Figure 1. **Conceptual model of human-supervisory control (modified from Pina et al. [5]).**

This paper presents a brief overview of existing generalizable metric frameworks and then defines 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. 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 explicitly measure human performance.

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. suggested identifying common metrics in terms of three aspects: human, robot, and the system [7]. 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 more comprehensive framework for human-automation team performance based on a high-level conceptual model of human supervisory control. Figure 1 represents this conceptual model for a team of two humans collaborating, with each controlling an autonomous platform. The platforms also collaborate autonomously. These collaboration layers are 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, human behavior precursors, and collaborative metrics (Table 1). Mission effectiveness includes the 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 initial state, including attitudes and cognitive constructs that can be the cause of and can 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 can result in a comprehensive understanding of the human-automation performance, covering issues ranging from automation capabilities to human cognitive abilities. However, there still is a lack of an objective methodology to select a collection of metrics that most efficiently measure a system's human-automation performance. The following section presents a preliminary list of evaluation criteria that can help researchers evaluate the quality of a set of metrics.

Table 1. Human supervisory control metric classes and subclasses [9]

1) Mission Effectiveness (e.g., key mission performance parameters)
2) Automation Behavior Efficiency (e.g., usability, adequacy, autonomy, reliability)
3) Human Behavior Efficiency
   a) Attention allocation efficiency (e.g., scan patterns, prioritization)
   b) Information processing efficiency (e.g., decision making)
4) Human Behavior Precursors
   a) Cognitive precursors (e.g., situational awareness, mental workload)
   b) Physiological precursors (e.g., physical comfort, fatigue)
5) Collaborative Metrics
   a) Human/automation collaboration (e.g., trust, mental models)
   b) Human/human collaboration (e.g., coordination efficiency, team mental model)
   c) Automation/automation collaboration (e.g., platforms' reaction time to situational events that require autonomous collaboration)

3. METRIC EVALUATION CRITERIA

The proposed metric evaluation criteria for human supervisory control systems consist of five general categories that are listed in Table 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 awareness), the associated measures (e.g., subjective responses), and the measuring techniques (e.g., questionnaires given at the end of experimentation).

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.

Table 2. Metric evaluation criteria

1) Experimental Constraints (e.g., time required to analyze a metric)
2) Comprehensive Understanding (e.g., causal relations with other metrics)
3) Construct Validity (e.g., power to discriminate between similar constructs)
4) Statistical Efficiency (e.g., effect size)
5) Measurement Technique Efficiency (e.g., intrusiveness to subjects)

3.1 Experimental Constraints

Time and monetary cost 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 temporary 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 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 can enable researchers to have high experimental control, and 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. The types of measures that can be collected are constrained by the testing environment. For example, the responses to rare events are more applicable for research conducted in simulated environments, whereas observational measures can provide better value in field testing.

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 not possible to collect all required metrics. Therefore, each metric should be evaluated based on how much it explains the phenomenon of interest and how much it helps explain the underlying reasons for what other metrics measure.

The most important aspect of a study is finding an answer to the primary research question. The proximity of a metric to answering the primary research question defines the importance of that metric. For example, a workload metric may not tell much without a mission effectiveness metric. However, this does not mean that the workload metric 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 already collected. For example, rather than having two metrics that measure mission effectiveness, having one metric that measures mission effectiveness and another metric that measures human behavior can provide a better understanding on the team 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 the underlying reasons to what the other metrics measure. For example, operator response to an event, hence human behavior, will often be dependent on the conditions and/or operator's state when the event occurs. The response to an event can be described in terms of three set of variables [11]: 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 for an event can be better understood if the initial conditions and operator's state when the event occurs is 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 and automation behavior in order to assess the operator state and environmental conditions which may influence human behavior.

3.3 Construct Validity

Construct validity refers to how well the associated measure captures the metric or construct of interest. For example, subjective measures for 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, 12]. However, self-ratings assess meta-comprehension rather than comprehension of the situation: it is unclear whether or not operators are aware of their lack of situational awareness. Therefore, subjective responses on situational awareness are not valid to assess the 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 [13]. For example, the primary task performance starts to break down only when the workload reaches higher levels [13, 14]. Therefore, primary task performance measures are not sensitive to changes in the workload at lower workload levels, since with sufficient spare processing capacity the 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 human workload 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., Levin et al. [15]). However, even if workload and stress are related, they still are two separate metrics. Therefore, galvanic skin response cannot alone 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 were repeatedly collected from the same subject under identical conditions.

Intra- and inter-subject reliability is especially of concern for subjective measures. For example, self-ratings are widely utilized for mental workload assessment [16, 17]. 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 not expensive. 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 [18], 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 & 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.

High correlation between different measures, even if they are intended to assess different metrics, is another limiting factor for metric selection. A high correlation can be indicative of the fact that multiple measures assess the same metric or the same phenomenon. Hence, including multiple measures that are highly correlated with each other can result in wasted resources.

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 that are correlated with each other would inflate 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) [19]. 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 is expected to produce differences, and excluding others that are not expected to show differences among many treatments.

Another metric characteristic that needs to be considered is the number 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 [20], 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 very rare, and operators can go through 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, small frequency of observed events 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 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. 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 control on the experiment, such as using sub-populations, can lead to less generalizable results, and there is a tradeoff between the two.

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 operator's visual attention [21, 22]. In particular, 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 operator's situational awareness [23]. 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 then assesses global situational awareness metric by calculating the accuracy of operator's responses. The collection of the measure requires the interruption of the task in a way that is unrepresentative of the reality generating an artificial setting. The interruption may also interfere with other metrics such as operator's performance and workload, as well as other temporal-based metrics.

4. DISCUSSION

Supervisory control of automation is a complex phenomenon with high levels of uncertainty, time-pressure, and a dynamically-changing environment. The performance of human-automation teams depend on multiple components such as human behavior, 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 paper presented a preliminary list of evaluation criteria for determining a set of metrics for a given research question. These criteria were populated under five major headings: 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 the 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 paper should be applied to a collection of metrics rather than each individual metric alone, taking the interactions between different metrics into consideration. The list of evaluation criteria presented in this paper is a guideline 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. Future research will identify a methodology based on a cost-benefit analysis approach, which will objectively identify the best set of metrics for classifications of research studies.

5. ACKNOWLEDGMENTS

This research is funded by the US Army Aberdeen Test Center.

6. REFERENCES

[1] Sheridan, T.B., *Telerobotics, automation, and human supervisory control*. 1992, Cambridge, MA: The MIT Press.

[2] Scholtz, J., et al., *Evaluation of human-robot interaction awareness in search and rescue*, in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. 2004: New Orleans.

[3] Cooke, N.J., et al., *Advances in measuring team cognition*, in *Team cognition: understanding the factors that drive process and performance*, E. Salas and S.M. Fiore, Editors. 2004, American Psychological Association: Washington, D. C. p. 83-106.

[4] Olsen, R.O. and M.A. Goodrich, *Metrics for evaluating human-robot interactions*, in *Proceedings of NIST Performance Metrics for Intelligent Systems Workshop*. 2003.

[5] Pina, P.E., et al., *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. 2008: Amsterdam, The Netherlands.

[6] Crandall, J.W. and M.L. Cummings, *Identifying predictive metrics for supervisory control of multiple robots.* IEEE Transactions on Robotics - Special Issue on Human-Robot Interaction, 2007. **23**(5): p. 942-951.

[7] Steinfeld, A., et al., *Common metrics for human-robot interaction*, in *Proceedings of the 1st Annual IEEE/ACM Conference on Human Robot Interaction (Salt Lake City, Utah)*. 2006, ACM Press: New York, NY.

[8] Wickens, C.D., et al., *An Introduction to Human Factors Engineering*. 2nd ed. 2004, Upper Saddle River, New Jersey: Pearson Education, Inc.

[9] Pina, P.E., B. Donmez, and M.L. Cummings, *Selecting metrics to evaluate human supervisory control applications*. 2008, MIT Humans and Automation Laboratory: Cambridge, MA.

[10] Vidulich, M.A. and E.R. Hughes, *Testing a subjective metric of situation awareness*, in *Proceedings of the Human Factors Society 35th Annual Meeting*. 1991, The Human Factors and Ergonomics Society: Santa Monica, CA. p. 1307-1311.

[11] Donmez, B., L. Boyle, and J.D. Lee, *The impact of distraction mitigation strategies on driving performance.* Human Factors, 2006. **48**(4): p. 785-804.

[12] Taylor, R.M., *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.

[13] Eggemeier, F.T., C.A. Shingledecker, and M.S. Crabtree, *Workload measurement in system design and evalution*, in *Proceeding of the Human Factors Society 29th Annual Meeting*. 1985: Baltimore, MD. p. 215-219.

[14] Eggemeier, F.T., 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*. 1983: Norfolk, VA. p. 139-143.

[15] Levin, S., et al., *Tracking workload in the emergency department.* Human Factors, 2006. **48**(3): p. 526-539.

[16] Wierwille, W.W. and J.G. Casali, *A validated rating scale for global mental workload measurement applications*, in *Proceedings of the Human Factors Society 27th Annual Meeting*. 1983: Santa Monica, CA. p. 129-133.

[17] Hart, S.G. and L.E. Staveland, *The subjective workload assessment technique: a scaling procedure for measuring mental workload*, in *Human Mental Workload*, P. Hancock and N. Meshkati, Editors. 1988, North Holland B. V.: Amsterdam, The Netherlands. p. 139-183.

[18] O'Donnell, R.D. 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, Editors. 1986, Wiley Interscience: New York. p. 42-1 - 42-49.

[19] Johnson, R.A. and D.W. Wichern, *Applied multivariate statistical analysis*. Fifth ed. 2002, NJ: Pearson Education.

[20] Sheridan, T.B., *Humans and automation: system design and research issues*. 2002, New York, NY: John Wiley & Sons Inc.

[21] Janzen, M.E. and K.J. Vicente, *Attention allocation within the abstraction hierarchy.* International Journal of Human-Computer Studies, 1998. **48**: p. 521-545.

[22] Donmez, B., L. Boyle, and J.D. Lee, *Safety implications of providing real-time feedback to distracted drivers.* Accident Analysis & Prevention, 2007. **39**(3): p. 581-590.

[23] Endsley, M.R., B. Bolte, and D.G. Jones, *Designing for situation awareness: an approach to user-centered design*. 2003, Boca Raton, FL: CRC Press, Taylor & Francis Group.