INTERNATIONAL JOURNAL OF INTELLIGENT CONTROL AND SYSTEMS
VOL.16,NO.2,JUNE 2011,67-78

Metric Selection for Evaluating Human Supervisory Control of Unmanned Vehicles

Birsen DONMEZ and M. L. CUMMINGS

Abstract- **While broad metric classes have been proposed in the literature to facilitate metric selection for evaluating human-unmanned vehicle (UV) interaction, there still lacks a systematic method for selecting an efficient set of metrics from the many metrics available. Through an experiment with subject matter experts, we investigated which metric characteristics human factors practitioners consider to be important in evaluating human supervisory control of UVs. We tested two different multi-criteria decision making methods to help practitioners assign subjective weights to metric evaluation cost/benefit criteria. The majority of participants rated the metric evaluation criteria used for both tools as very useful. However, the majority of participants' metric selections before using the methods were the same as the suggestions provided by the methods. Since determining weights of cost/benefit importance is an inherently subjective process, the real value of using even objective computational tools may be reminding human factors practitioners of the important experimental criteria and relationships between these criteria that should be considered when designing an experiment.**

***Index Terms*—Metrics, metric quality, human supervisory control, AHP, human factors, measurement techniques, and experiments**

1. INTRODUCTION

Human-automation teams are common in many domains, such as command and control operations, human-robot interaction, process control, and medicine. With high levels of 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 [1]. Operation of unmanned vehicles (UVs) is one particular domain where humans are increasingly placed in a supervisory role. The role of the human operator has changed from attending to low-level tasks (e.g., manually flying an aircraft) to higher-level tasks such as monitoring and generating contingency courses of action.

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, or the human-UV team in the context of UV operations. However, this metric fails to provide insights into the process that leads to the final mission-related output. Measuring multiple human-automation team aspects, such as workload and usability 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, whether the experiment is intended to compare different conditions in a laboratory setting or to evaluate the system performance against some objective for more real-world applications.

Many practitioners select their metrics based on 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 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 an efficient set of metrics among the large number of available metrics.

Although metric selection issues apply broadly to human performance research questions, this research focuses specifically on human supervision of UVs to identify the "quality" of common/applicable metrics, and explore the usefulness of generic multi criteria decision making methods for UV metric selection.

In this paper, we first summarize earlier efforts on the development of metric evaluation criteria based on a comprehensive review of metrics applicable to human supervision of UVs. We then present an experiment conducted with human factors practitioners to explore the usefulness of two existing multi criteria decision making methods designed to assist practitioners in the metric selection process using the evaluation criteria. We also report subject matter expert opinions on the usefulness of the proposed metric evaluation criteria. We conclude with recommendations for future applications.

2. METRIC EVALUATION CRITERIA

Different frameworks of metric classes are found in the literature in terms of human-UV interaction [4-7]. These frameworks categorize existing metrics into high-level metric classes that assess different aspects of the human-automation performance and are generalizable across missions. Pina et al. [5] defined five generalizable metric classes for supervisory control of UVs: mission effectiveness, automation behavior efficiency, human behavior efficiency, human behavior precursors, and collaborative metrics. These metric classes can help experimenters select metrics that result in a comprehensive

Manuscript received Dec. 10, 2010, revised Feb. 25, 2011, accepted Mar 31, 2011. This work was supported by US Army Aberdeen Test Center. This paper is extended from "Metric selection for evaluating human supervisory control of unmanned vehicles" published at *Performance Metrics for Intelligent Systems Workshop,* Baltimore, MD, USA, Sept., 2010.

Birsen Donmez is with Dept. of Mechanical and Industrial Engineering, University of Toronto, 5 King's College Rd, Toronto, ON, M5S 3G8, Canada (e-mail: donmez@mie.utoronto.ca); M. L. Cummings is with Dept. of Aeronautics and Astronautics, Massachusetts Institute of Technology, 77 Mass Ave., Cambridge, MA, 02139, USA (e-mail: missyc@mit.edu).

understanding of the human-UV performance, covering issues ranging from automation capabilities to human cognitive abilities. For holistic system assessment, a rule of thumb is to select at least one metric from each metric class. However, there is a lack of a systematic methodology to select a collection of metrics across these classes. Each metric set has advantages, limitations, and costs, thus the added value of different metric sets for a given context needs to be assessed in order to select an efficient set that maximizes value and minimizes cost.

Donmez, Pina, and Cummings [8] proposed a list of metric evaluation criteria for human supervisory control of UVs: experimental constraints, comprehensive understanding, construct validity, statistical efficiency, and measurement technique efficiency. This list was identified through a comprehensive literature review of different metrics and measuring techniques applicable to UV supervisory control [9]. The following section briefly presents these criteria. Detailed discussions and UV supervisory control metric examples can be found in [8, 10].

It should be noted that the costs and benefits of different research techniques in human engineering have been previously discussed in the literature [11, 12]. For example, Bethel and Murphy [13] present a useful discussion on the advantages and disadvantages of methods that can be utilized specifically to evaluate human-robot interaction. However, the dimension they use in their discussion is data collection method (e.g., self assessment, interviews, etc.) rather than metric quality. Our list of metric evaluation criteria focuses specifically on metric quality and inherently encompasses information on different data collection methods.

2.1 Experimental Constraints

Time and monetary cost associated with measuring and analyzing a specific metric constitute the main practical considerations for metric selection. Availability of temporal and monetary resources depends on the individual project and such factors are typically limiting in all projects. The stage of system development and the testing environment are additional constraints that can guide metric selection. 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. In general, 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.

2.2 Comprehensive Understanding

It is important to maximize the understanding gained from a research study. Given that it is often not possible to collect all required metrics, each metric should be evaluated based on how much it explains the phenomenon of interest or its coverage. For example, continuous measures of workload over time (e.g., pupil dilation) can provide a more comprehensive dynamic understanding of one aspect of a 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 or evaluation question. The proximity of a metric to answering this 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. 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, a workload measure can provide additional insights beyond just human-UV 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, the underlying reasons for an 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 [5] in order to assess the operator state and environmental conditions, which may influence human behavior.

2.3 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 participants to rate the amount of situational awareness they had on a given scenario or task. These measures are proposed to help in understanding participants' situational awareness [14, 15]. 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.

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 [16]. For example, primary task performance generally starts to break down when the workload reaches higher levels [16, 17], thus primary task performance measures are not sensitive to changes at lower workload levels.

A measure with high construct validity should also be able to discriminate between similar constructs. An example measure that fails to discriminate two related

metrics is galvanic skin response, which has been proposed and used to measure workload and stress levels (e.g., [18]). However, even if workload and stress are related, they are still 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 participant, whereas intra-subject reliability requires the measure to assess the same construct if the measure were repeatedly collected from the same participant under identical conditions. For example, self-ratings are widely utilized for mental workload assessment [19, 20]. However, different individuals may have different interpretations of workload, leading to decreased inter-subject reliability. Some participants may not be able to separate mental workload from physical workload [21], and some participants may report their peak workload, whereas others may report their average workload. 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.

2.4 Statistical Efficiency

Three metric qualities 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) [22]. 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 [23].

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, thus human monitoring efficiency is an important metric to measure. 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, observed events with a low frequency of occurrence cannot be statistically analyzed unless data is obtained from a very large number of participants, such as in medical studies on rare diseases.

The number of participants 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. One way to compensate for limited number of participants in a study is to use more sensitive measures that will provide a large separation between different conditions, that is, a high effect size.

2.5 Measurement Technique Efficiency

The data collection technique associated with a specific metric should not be intrusive to the participants or to the nature of the task. For example, eye trackers can be used for capturing operators' visual attention (e.g., [24, 25]). However, head-mounted eye trackers can be uncomfortable for participants, and hence influence their responses. Wearing an eye-tracker can also lead to an unrealistic situation that is not representative of real world tasks. Thus, the measurement technique can interfere with the construct validity of a metric.

The measuring technique can also interfere with the realism of the study. For example, off-line query methods are used to measure operators' situational awareness [26], by briefly halting the experiment at randomly selected intervals, blanking the displays, and administering a battery of queries to the operators. 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.

3. Multi Criteria Decision Making Methods for Metric Selection

Donmez, Pina, and Cummings [8] translated the above criteria into potential cost-benefit parameters (Table 1), which can be used to define cost and benefit functions of a metric set, Eqn. (1). Some criteria can be considered as a benefit or a cost (e.g., non-intrusiveness vs. intrusiveness). Thus, the breakdown in Table 1 was based on the ability to assign a monetary cost to an item.

$$Benefit\ of\ metric\ I = \sum_{i=1}^{NB} WB_i \times MB_{Ii} \quad \text{where}$$

WB_i: weight of importance for benefit criterion i
MB_{Ii}: how well metric I meets benefit criterion i
NB: total number of benefit criteria

$$Cost\ of\ metric\ I = \sum_{j=1}^{NC} WC_j \times MC_{Ij} \quad \text{where} \tag{1}$$

WC_j: weight of importance for cost criterion j
MC_{Ij}: how much metric I costs for cost criterion j
NC: total number of cost criteria

Table 1: An example breakdown of 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 Institutional Review Board 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		

Depending on research objectives and limitations, the entries in the cost and benefit functions can have different weights of importance (i.e., WB_i and WC_j). Two promising techniques identified to help researchers assign subjective weights are the pair-wise comparison approach of the analytic hierarchy process (AHP) [27], and the ranking approach of the probability and ranking input matrix (PRIM) method [28]. Direct assignment of weights is not adopted as an alternative since humans have difficulty with absolute judgment and are better at making relative judgments [12].

AHP is widely used both in academic research and in the industry. It begins with the user building a decision hierarchy, which includes the goals (e.g., identify metric benefits), decision alternatives (e.g., NASA TLX, pupil dilation), and criteria (e.g., non-intrusiveness, construct validity). There are no systematic guidelines for creating the hierarchy or identifying the decision alternatives and criteria. The hierarchies depend on user knowledge and experience.

At each level of a hierarchy, AHP utilizes pair-wise comparisons to express the relative importance of one criterion over another (e.g., Fig. 1). The relative importance is judged on a five point Likert scale with the end values of equally important and extremely more important. The values obtained from pair-wise comparisons are then used to create a weight matrix. The eigenvectors of this weight matrix correspond to the criteria weights of interest. There are disadvantages associated with AHP identified in the literature suggesting flaws in the methods of combining individual weights into composite weights [29, 30].

Another characteristic of AHP, potentially a user acceptance issue, is the consistency checks that are imposed on the user. AHP forces the user to perform all possible pairwise comparisons even if some of these comparisons are redundant. For example, if the user is comparing A, B, and C, then a comparison between A and B and a comparison between B and C would indicate how A and C would compare. Even if a comparison of A and C is redundant, AHP forces the user to perform it until a consistency criterion is met (consistency ratio$\leq$ 0.1 as suggested by Saaty [31]), with the claim that consistency checks help the user think about his ratings in detail. The consistency ratio criterion of 0.1 is an arbitrary cutoff but is the convention.

Fig. 1. Interface for Analytic Hierarchy Process (AHP) used in the metric selection experiment (Section 4).

The consistency ratio takes into account not only the directionality of the responses but also the magnitude. For example, when comparing A, B, and C, if the user indicates that both A and B are moderately more important than C, then he has to indicate that A and B are equally important. Rating A to be even slightly more important than B (or vice versa) would lead to a consistency ratio of 0.19 and would be considered incorrect by AHP. Thus, AHP does not always allow for finer grain comparisons.

The ranking input matrix (RIM) is similar to more traditional engineering decision matrices such as the ones used in quality function deployment [32]. The RIM method allows people to categorically select weights through a direct perception-interaction interface (e.g., Fig. 2) [28]. Each item is represented by a puck that can slide (through clicking and dragging) onto a ranking matrix.

The ranking matrix consists of 10 slots consisting of five main categories of importance: high, medium-high, medium, low-medium, and low. Each of these main categories has two bins to allow the person to indicate slight variations in the importance of items. The pucks can also be placed side by side indicating equal importance. A numeric weight value is assigned to these bins on a scale of 0.05 to 0.95 with 0.10 intervals.

AHP creates hierarchies and only the entries in one level of a hierarchy are directly compared by the user. In contrast, RIM allows the users to see the weights in each category side by side, and manipulate them if necessary. In general, AHP is not as transparent and thus may be harder for the decision makers to understand. Moreover, the inability to directly compare subcategories in a hierarchy can generate unexpected results.

In addition to requiring subjective weights of importance, the cost and benefit functions (1) also require values representing how well each metric meets the evaluation criteria (i.e., MB_{Ii} and MC_{Ij}). In some cases, the value of a metric can be represented with an objective number (e.g., time required to collect a metric). However for many criteria, finding an objective value is impossible (e.g., construct validity of a metric).

In addition to determining WB_i and WC_j, AHP and RIM can also be used to determine subjective values for MB_{Ii} and MC_{Ij}. In application, a user would conduct either pairwise comparisons (for AHP) or rankings (for RIM). The methods then would generate values for weights of importance for the evaluation criteria (WB_i and WC_j) and how well each metric meets each criterion (MB_{Ii} and MC_{Ij}). The user would then be presented with cost and benefit values calculated via Eqn. (1) or a combination of them (e.g., difference or ratio of cost and benefit values).

Both AHP and RIM are intended to help decision makers select a choice out of many. However, when trying to answer a research question, researchers will most likely need more than one metric. When selecting multiple metrics, the benefits and costs for multiple metrics will need to be combined. Moreover, the dependencies between the selected metrics will also need to be incorporated into the combined benefit-cost. For example, the total number of metrics selected would have an influence on the type I error of each individual metric.

The linear combination of benefit-cost values facilitates both the combination of multiple metric costs and benefits, as well as the incorporation of metric dependencies by allowing additional terms to be added or subtracted from the overall value. Therefore, in the experiment described in the next section, we used the difference of benefit and cost values to rank the metrics, but also presented the corresponding benefit and cost values to the participants. Another approach is to mimic how humans would combine these cost-benefit values.

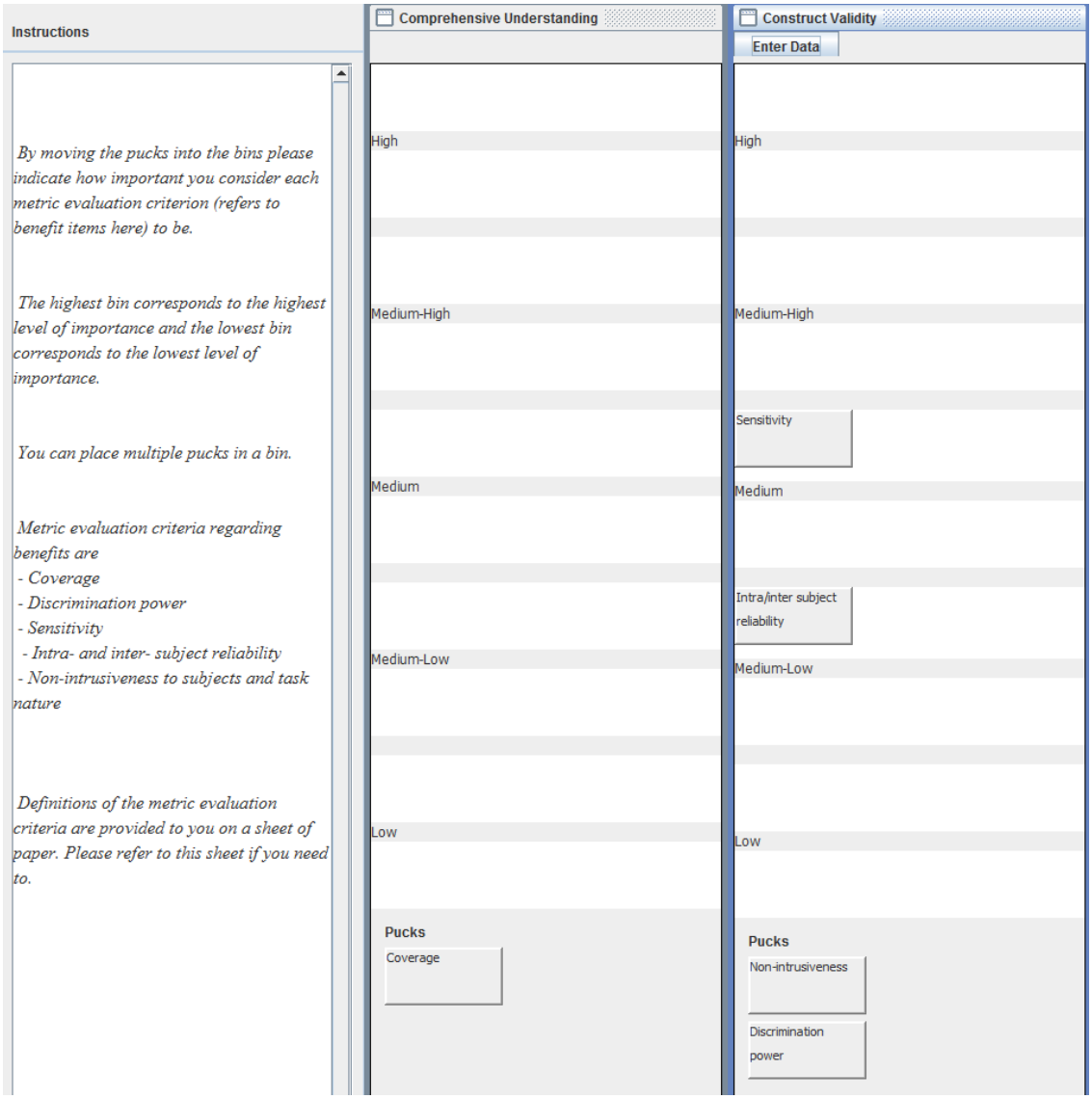

Fig. 2. Interface for Ranking Input Matrix (RIM) used in the metric selection experiment (Section 4).

Our current approach may not be optimal, however, the best method, if one exists, is currently unknown and is an area for future research. However, given that selection of multiple metrics is more realistic than selecting a single metric, it is important to facilitate the incorporation of metric dependencies when combining benefit and cost values. It is also important to assess if people can account for metric dependencies (e.g., statistical implications of collecting multiple metrics) when they evaluate metrics against a set of criteria. The latter issue was investigated as part of a larger experiment conducted to evaluate AHP and RIM methods for metric selection.

4. METRIC SELECTION EXPERIMENT

An experiment was conducted to a) investigate the perceived usefulness of the metric evaluation criteria, b) identify which criteria human factors researchers consider to be important, and c) evaluate AHP and RIM for supporting metric selection. Although we had some expectations based on the characteristics of AHP and RIM (e.g., time to complete AHP would be much longer), we did not set any directional hypothesis prior to this experiment. We wanted to assess the insights of subject matter experts on the evaluation criteria as well as potential usefulness of AHP and RIM in metric selection.

Thirty-one human factors practitioners were presented with the description of a hypothetical UV supervisory control experiment, which was adapted from an actual experiment conducted by Donmez, Cummings, and Graham [33]. The participants were then asked to select either one or multiple workload metrics for this hypothetical experiment from a list of potential workload metrics provided to them. After making an initial selection,

the participants used both AHP and RIM (order counterbalanced) to evaluate the list of workload metrics. After AHP and RIM solutions were displayed, the participants were given the choice to change their initial metric selection. They could keep their initial selection, pick AHP or RIM solutions, or come up with an entirely different selection. At the end of the experiment, the participants filled out a questionnaire, evaluating AHP and RIM on a multitude of characteristics.

In order to prevent experimental confounds, we focused only on workload metrics. Workload is a common human factors metric that most human factors practitioners understand [34, 35], and is a common metric gathered across human-UV studies. Participants were not allowed to select a workload metric that was not on the list provided to them. Keeping the experiment bounded allowed for a shorter experiment and more control on the experimental conditions, hence a better ability to draw conclusions.

4.1 Participants

A total of 31 participants completed the study, and all had experience with human subject experimentation and metrics. Experience with human subject experimentation ranged from one month to forty years. Participants were recruited from both academia and industry, and consisted of 9 females and 21 males, ages ranging from 19 to 64 years (average: 36.6, stdev: 13.6). Eleven of the participants currently held an academic position. The highest degrees held included high school (n=1), college (n=12, 5 in academia and 7 in industry), Master's (n=12, 4 in academia and 8 in industry), and Ph.D. (n=6, 2 in academia and 4 in industry). The experiment took 1 to 1.5 hours to complete.

4.2 Apparatus

The experiments were conducted in a mobile experimental test-bed mounted in a 2006 Dodge Sprinter. Two 21-inch wall mounted displays were used in the experiment. By integrating an experimental test bed into a vehicle, the experiment was able to travel to the participants. Access restrictions into government facilities, particularly with foreign graduate students, often make it difficult to take such experiments directly into the work place. Thus, the use of the vehicle allowed a high number of human factors practitioners to be recruited for participation.

4.3 Experimental Design

The experiment was a 2x2 mixed factorial design with two independent variables: number of metrics to select (a single metric, a subset of all metrics) and weight assignment method (AHP, RIM). Number of metrics to select was a between-subjects variable, with 15 participants selecting a single metric out of all the candidate metrics, and another 16 selecting a subset of all the metrics (one, two, or all). Weight assignment technique was a within-subjects variable with each participant making a decision using both AHP and RIM. In order to control for learning effects, the order of presentation was counterbalanced.

4.4 Experimental Tasks

The experimental instructions started with the description of the hypothetical experiment and the list of potential workload metrics to choose from: embedded secondary task performance, NASA TLX, and pupil dilation based on eye tracking data. The hypothetical experiment assessed the effects of different auditory alerts on human supervision of multiple unmanned aerial vehicles. When participants finished reading this part of the instructions, they were asked to select either one or a subset of workload metrics depending on the experimental condition they were assigned (i.e., number of metrics to select).

After the initial metric selection, participants read a detailed description of the metric evaluation criteria. A subset of the criteria identified in Donmez, Pina, and Cummings [8] was selected to be included in this experiment. The selection was based on the relevance to the metrics used in the hypothetical experiment. The cost estimates were provided where applicable. There were no explicit monetary or time constraints imposed on the experiment. To have more experimental control, we did not ask the participants to define a hierarchy structure for AHP but provided the structure below.

Benefits:

- Coverage
- Construct validity:
  - a) discrimination power
  - b) sensitivity
  - c) inter/intra subject reliability
  - d) non-intrusiveness
- Type I error (for multiple metric selection)

Costs:

- Data gathering:
  - a) time for data collection
  - b) monetary cost for data collection
  - c) measurement error likelihood
- Data analysis:
  - a) time for analysis
  - b) expertise for analysis

The instructions included a detailed description of AHP and RIM, including how the benefit-cost values were calculated. After reading about the first method (AHP or RIM) the participants used an interface for that method. With this interface, the participants assigned subjective weights of importance to the metric evaluation criteria, and also determined how well potential workload metrics met each criterion. In the RIM condition, the participants used

the click and drag interfaces (Fig. 2) to rank the evaluation criteria based on importance, as well as to rank the metrics with respect to how well they met the criteria. In the AHP condition, participants conducted pair-wise comparisons to indicate the relative importance of evaluation criteria, and within each criterion they performed pair-wise comparisons to identify how well the metrics satisfied the criteria (Fig. 1). Instructions were also provided on the interfaces as reminders on what to do for each window. Since the complete set of written instructions was available throughout the experiment, the participants could also refer back to them if they needed clarification.

In AHP, if participants could not meet the consistency threshold of 0.1 suggested by Saaty [31], then they were presented with a pop-up window indicating their inconsistency. The participants were asked to retry and change their responses to achieve the suggested consistency threshold. However, participants were given the ability to skip this step if they felt they had tried "many" times but could not reach the threshold value. The ability to skip was deemed important since we observed in pilot testing that participants would get frustrated to the point that they wanted to quit the experiment. The details on consistency checks were included in the written instructions and were also demonstrated to the participants before they started the AHP trial.

After completing the session with the first interface, the participants read the instructions for the next method (AHP or RIM) and completed their second test session using the next interface.

The experimental tasks for the multiple metric selection condition were slightly different than the single metric selection condition. As previously mentioned, the participants in this condition were told that they could select more than one metric. These participants were also presented with an extra evaluation criterion: type I error. This criterion is not relevant for single metric selection, however, it can be a negative benefit when selecting multiple metrics since analyzing more metrics increases the overall type I error. Participants compared this criterion to the other criteria in terms of importance. In order to assess if participants were aware of how much type I error would change with different number of metrics, they were also asked to compare the number of workload metrics collected (1 to 3) with respect to type I error.

At the end of the experiment, participants were provided with the suggested list of workload metrics ranked based on AHP or RIM solutions. In the multiple metric selection condition, this list could consist of groupings of metrics. For example, the best solution could be NASA TLX and secondary task performance. The participants were then asked to evaluate the solutions provided by AHP and RIM and the initial selection they indicated before using the interfaces. This evaluation helped us assess if the two methodologies result in different selections and if so, which methodology produces results regarded to be better by the participants. Post-test surveys were administered to assess participant opinions about the evaluation criteria and the two methods.

5. EXPERIMENTAL RESULTS

Based on modeling assumptions, mixed linear models were built for continuous data, whereas non-parametric statistics were utilized to analyze categorical data where appropriate (α=.05).

5.1 Selected Metrics

For single metric selection, AHP and RIM in general resulted in the same solutions (87%), which also matched most of the participants' initial choices (AHP: 73%, RIM: 87%). Thus, regardless of the method used, participants directed each tool so that the results generally matched their expectations. Participants' self reported experience with the three workload metrics was assessed on a Likert scale (1: no experience, 5: expert). Participants in general had more experience with secondary task (mean=2.3) and NASA TLX (mean=2.3) measures as compared to pupil dilation (mean=1.8). There were approximately an equal number of participants (n=8) who identified secondary task and/or NASA TLX as the metric they have the most experience with. Regardless of this previous experience, 10 out of 15 participants still chose secondary task as their initial metric selection rather than NASA TLX, suggesting that previous experience did not solely determine the selected metric.

For multiple metric selection, the majority (n=9) of the participants selected secondary task and NASA TLX as their preferred metrics, which was followed by NASA TLX (n=3) as the second most preferred metric. Interestingly, contrary to our expectation, many of the participants did not choose to collect as many metrics as they could. This finding may be due to the experimental instructions that highlighted resource limitations. Similar to the single metric selection condition, there was no strong evidence to suggest that the participants changed their selections based on the advice from one or the other method.

5.2 Benefit Criteria Weights

The single metric selection condition had five benefit criteria (coverage, discrimination power, sensitivity, inter- and intra- subject reliability, and non-intrusiveness), whereas the multiple metric selection condition had an additional criterion (type I error), bringing the total to six.

In the single metric selection condition, differences in benefit weights were observed for both RIM and AHP. For RIM, discrimination power and coverage received significantly higher weights compared to sensitivity, inter- and intra- subject reliability, and non-intrusiveness ($F(4, 56)=3.22$, $p=.02$). For AHP, coverage had a significantly higher weight than all other benefit criteria ($F(4, 56)=7.99$, $p<.0001$).

Differences in benefit criteria weights were also observed in the multiple metric selection condition. For AHP, similar to the single metric selection condition, coverage resulted in a higher weight than all other benefit criteria ($F(5, 75)=21.71$, $p<.0001$). For RIM, although coverage and discrimination power had the highest average weight estimates, the statistical analysis did not reveal significant results ($F(5, 75)=1.78$, $p=.13$).

To summarize, participants generally considered coverage and discrimination power as the most important among other benefit criteria. Therefore, if a metric was considered to have high coverage or discrimination power, it was more likely to be preferred. The underlying theoretical reasons for this weighting scheme are unclear and this area deserves further research focus.

5.3 Type I Error

In this experiment, we focused on type I error as a way of assessing if researchers think about the more hidden ramifications of collecting multiple metrics aside from monetary or time costs.

In the multiple metric selection condition, as part of RIM and AHP, participants were asked to rate how having one, two, or three metrics would affect overall resulting type I error. Six participants out of the 16 total incorrectly indicated that either the overall type I error would not be impacted (n=1) or the type I error would increase as the number of metrics decrease (n=5). Three of these six participants repeated their mistake twice, once with RIM and once with AHP. There were no particular common characteristics for the participants who repeated their mistake. It is unclear if the incorrect responses regarding type I error were due to slips or mistakes. That is, they could be either due to a failure to follow the interface instructions or a lack of knowledge. Regardless of the cause, a fallacy of both methods is that the outputs from AHP and RIM are only as good as the information provided to them.

5.4 Subjective Ratings

The evaluation criteria received an average usefulness rating of 4.4 (1-lowest, 5-highest). There was one response with a rating of 3, 18 responses of 4, and 12 responses of 5.

Participants were also asked a list of 1-5 Likert scale questions to assess their understanding and perceived usefulness for the two methods. Table 2 presents statistical results comparing participant ratings with respect to being less than or equal to average vs. being above average (χ^2). Overall, participants' ratings for RIM indicated greater than average perceived usefulness, understandability, and worthiness of their time. For AHP, these responses were not significant, except a marginally significant result assigned to understandability.

Table 2: Subjective ratings on method usefulness, understanding (* significant at $\alpha=.05$)

		1 Low	**2**	**3 Avg.**	**4**	**5 High**	**χ^2 (p-value) (4-5 vs. 1-3)**
Usefulness	AHP	0	6	7	10	8	.81 (.47)
	RIM	0	3	5	17	6	7.26 (.01)*
Worth the time	AHP	1	6	6	15	3	.81 (.47)
	RIM	0	2	6	20	3	7.25 (.01)*
Understand method	AHP	2	1	7	10	11	3.9 (.07)
	RIM	0	1	8	8	14	5.45 (.03)*

5.5 Time for Metric Selection

Significant differences were observed on how long it took the participants to select their metric(s). AHP took on average 435 sec longer than RIM (95% CI: 307, 562), a 73% increase. There was no interaction effect, hence regardless of the method used, the second trial took on average 214 sec shorter than the first trial (95% CI: 127, 301), a 23% decrease. This finding was expected since both conditions used the same scenario.

5.6 AHP Consistency Conformance

Consistency was only an issue when evaluating three or more elements through pairwise comparisons. On average, participants were prompted to retry on 48% of such instances (stdev=20%). On average, the maximum number of times they had to retry in a single instance was 4.8 (stdev=3.2, min=1, max=14).

When the participants were prompted to retry at least once, they skipped without achieving the suggested consistency threshold on average 38% of the time (stdev=39%). Out of the 31 total participants, 11 retried until they achieved consistency (0% skip), whereas 5 chose to skip 100% of the time either after some retrials or none. The rest skipped occasionally with skip rates ranging from 8% to 86%. The skipping consistency values were on average 0.22 (stdev=0.13, max=0.65). The participants who skipped 100% of the time had an average age of 49, whereas the participants who tried until they reached consistency were younger (average age: 29). Experience with workload metrics were similar across the two groups ($t(14)=0.27$, $p=.8$).

5.7 Open-ended Comments on AHP and RIM

At the end of the experiment, participants were specifically asked to write down positive and negative aspects they identified with AHP and RIM.

The majority of the positive AHP comments addressed the pairwise comparisons (n=12 or 40% of participants). Thirteen percent of participants indicated that AHP made

them think longer and in more detail (n=4). Twenty three percent liked consistency checks (n=7), whereas 16% (n=5) identified them to be frustrating. Thus, the views on consistency checks were split. Thirty percent thought that AHP was too complicated (n=11), and 16% identified it as time consuming (n=5).

The positive aspects of RIM cited commonly were ease of use (n=10 or 32% of participants), ease of visualizing responses (n=9 or 29% of participants), speed (n=8 or 26% of participants), and simplicity (n=5 or 16% of participants). The total number of negative responses for RIM (n=11) was fewer than the total number of negative responses for AHP (n=32). A few participants indicated that they did not think critically at times (n=3 or 10% of participants). The 10-point rating scale was deemed hard by a few participants (n=3 or 10% of participants).

6. DISCUSSION

This paper presents an approach for helping experimenters select an efficient set of metrics for evaluating UV supervisory control. The metric evaluation criteria and the relevant cost-benefit parameters presented are guidelines only. It should be noted that there is not a single set of metrics that are efficient across all applications. Research-specific aspects such as available resources and the questions of interest will ultimately determine the relative metric quality.

Two different methods to develop principled subjective weights were identified and evaluated through an experiment with human factors practitioners: AHP and RIM. In order to keep the experiments short, participants were asked to evaluate only three workload metrics. Overall, the participants rated RIM to be more useful, easier to understand, and worth their time. AHP took a significantly longer time, and some participants considered it to be time consuming. In reality, researchers not only have to choose from a large number of metrics but they also ideally have to choose from a large number of constructs (e.g., performance, workload, etc.). Because AHP requires pairwise comparisons between all potential metrics, each additional potential metric would drastically increase the time required to perform AHP. Thus, the appropriateness of AHP selecting from a large set of potential metrics is questionable.

Another AHP problem revealed from the experiment is user frustration and/or lack of conformance to consistency checks. All participants experienced consistency issues where they could not meet the consistency threshold suggested by the AHP inventor, Saaty [31]. Some participants skipped achieving consistency 100% of the time, whereas some retried until they achieved the threshold. The participants who tried to achieve the threshold indicated that at times they forgot about what they were evaluating, and instead focused on tweaking their responses to obtain a value less than 0.1. In addition, some participants indicated that pairwise comparisons made them lose the big picture. These issues are potential concerns with any method that utilizes pairwise comparisons for assessing subjective responses (e.g., NASA TLX).

When it came to the metrics selected, the majority of participants' initial metric selections matched the solutions proposed by AHP and/or RIM. Thus, no substantial benefits were observed for either of the methods. Even if these methods use mathematical formulas to obtain cost benefit functions, they are inherently subjective as users provide most of the information that goes in the cost benefit functions. Therefore, if the user inputs incorrect information, either by a slip or a mistake, the methods may provide flawed results.

For example, participants were asked to indicate the effects of additional metrics on the overall type I error. Responses from 37% erroneously suggested that type I error decreases with additional metrics analyzed. Combined with the weight of importance for type I error, this erroneous information was included in AHP and RIM calculations. But because type I error was only one criterion among many and its weight of importance was not very high, the final solutions of AHP and RIM were not drastically influenced by the incorrect inputs.

While using AHP and RIM, participants referred back to the criteria several times as observed by the experimenter. Approaches like AHP and RIM have the potential to help researchers select metrics by considering many attributes that they may not consider otherwise. Thus, it is essential to provide better information to researchers in terms of how they could view the costs and benefits of a specific metric, before providing them with a mathematical tool that predicts what the best set of metrics would be. However, a mathematical tool might still prove to be useful when selecting metrics from various alternatives. The experiment presented in this paper focused on selecting from three alternatives only. With more possible alternatives in more complicated experiments, decision making becomes more complex, potentially warranting a decision support tool.

When selecting from a few workload metrics, time to complete AHP was reasonable, but RIM was much faster to use. Thus, for evaluating a larger set of metrics and more metrics of different types, RIM appears to be more appropriate. However, the acceptance and effectiveness of RIM for evaluating a larger set of metrics is currently unclear and should be investigated in the future. Future experiments with subject matter experts can evaluate the effectiveness and acceptance of RIM for the same hypothetical experiment used in the current study but with a larger metric set to choose from (e.g., performance and workload). Moreover, the underlying methodology for RIM can be modified in order to support metric selection when evaluating metrics from multiple classes. For example, a penalty can be introduced to avoid selecting metrics from the same class rather than selecting metrics from different classes. Determining such modifications in the RIM methodology is another point for future research. An experiment comparing RIM with and without such a

modification can provide insights into the effectiveness of the modification.

7. CONCLUSION

Recent dramatic advances in the unmanned vehicle domain are not just limited to military operations as the international civil sector is now looking to such unmanned technologies to aid operations such as fighting forest fires, undersea exploration, monitoring wildlife, inspecting bridges, and supporting first responders such as police and rescue. UAV expenditures alone are predicted to more than double in the next ten years, and are expected to exceed $80 billion [36].

Accompanying such rapid technological advancements is the need to evaluate not just the technology, but also the human-automation interaction. Without principled evaluation approaches to what fundamentally is an interdisciplinary system of systems endeavor, resulting technologies could fall short of expectations or potentially cause significant setbacks.

Towards this concept of principled evaluation, we identified five evaluation criteria (experimental constraints, comprehensive understanding, construct validity, statistical efficiency, and measurement technique efficiency) that should be considered when evaluating such unmanned systems. However, while this paper focused on human supervisory control of UVs (including human-robot interaction), our approach and many of the findings and recommendations apply more broadly to human performance evaluation. While these evaluation criteria may not be comprehensive for all domains, they can still be used as a guideline for assessing a metric's quality.

ACKNOWLEDGMENTS

Special thanks are extended to MIT undergraduate students, Meghan Dow, Grace Taylor, and Barden Cleeland, who helped with the experiment. Thanks to Brooke Abounader (US Army Aberdeen Test Center) and Luca F. Bertucelli (MIT) for their insightful comments.

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, *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)*, New York, NY: ACM Press, 2006.

[8] B. Donmez, P. E. Pina, and M. L. Cummings, "Evaluation criteria for human-automation performance metrics," in *Performance Evaluation and Benchmarking of Intelligent Systems*, R. Madhavan, *et al.*, Eds.: Springer, 2009, pp. 21-40.

[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] B. Donmez and M. L. Cummings, "Metric selection for human supervisory control," Humans and Automation Laboratory, Cambridge, MA HAL2009-05, 2009.

[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] C. L. Bethel and R. R. Murphy, "Review of human studies methods in HRI and recommendations," *International Journal of Social Robotics*, vol. 2, pp. 347-359, 2010.

[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] 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.

[16] F. T. Eggemeier, C. A. Shingledecker, and M. S. Crabtree, "Workload measurement in system design and evaluation," in *Proceedings of the Human Factors Society 29th Annual Meeting*, Baltimore, MD, 1985, pp. 215-219.

[17] 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.

[18] S. Levin, *et al.*, "Tracking workload in the emergency department," *Human Factors*, vol. 48, pp. 526-539, 2006.

[19] 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.

[20] 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.

[21] 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, *et al.*, Eds., New York: Wiley Interscience, 1986, pp. 42-1 - 42-49.

[22] R. A. Johnson and D. W. Wichern, *Applied Multivariate Statistical Analysis*, Fifth ed. NJ: Pearson Education, 2002.

[23] G. W. Oehlert, *A First Course in Design and Analysis of Experiments*. New York: W. H. Freeman and Company, 2000.

[24] 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.

[25] 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.

[26] 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.

[27] T. L. Saaty, *Fundamentals of Decision Making and Priority Theory with the Analytic Hierarchy Process*, 2nd ed. Pittsburgh, PA: RWS Publications, 2006.

[28] H. D. Graham, G. Coppin, and M. L. Cummings, "The PR matrix: extracting expert knowledge for aiding in C2 sense and decision making," in *Proceedings of the 12th International Command and Control Research and Technology Symposium*, Newport, RI, 2007.
[29] S. Schenkerman, "Inducement of nonexistent order by the anlaytic hierarchy process," *Decision Sciences*, vol. 28, pp. 475-482, 1997.
[30] R. D. Holder, "Some comments on the analytic hierarchy process," *The Journal of Operational Research Society*, vol. 41, pp. 1073-1076, 1990.
[31] T. L. Saaty, *The Analytic Hierarchy Process*. New York: McGraw-Hill, 1980.
[32] Y. Akao, *Quality Function Deployment: Integrating Customer Requirements into Product Design*. Cambridge, MA: Productivity Press, 1990.
[33] B. Donmez, M. L. Cummings, and H. D. Graham, "Auditory decision aiding in supervisory control of multiple unmanned aerial vehicles," *Human Factors*, vol. 51, pp. 718-729, 2009.
[34] V. J. Gawron, *Human Performance, Workload, and Situational Awareness Measures Handbook*, 2nd ed. Boca Raton, FL: CRC Press, 2008.
[35] S. G. Hill, *et al.*, "Comparison of four subjective workload rating scales," *Human Factors*, vol. 34, pp. 429-439, 1992.
[36] Teal-Group, "World Unmanned Aerial Vehicle Systems: 2010 Market Profile and Forecast," Teal Group Corporation, Manassas, VA, 2009.

Birsen Donmez received the B.S. degree in mechanical engineering from Bogazici University, Istanbul, Turkey, in 2001, the M.S. degree in industrial engineering in 2004, the Ph.D. degree in industrial engineering in 2007, and the M.S. degree in statistics in 2007, all from the University of Iowa, Iowa City.

She is currently an Assistant Professor with the Department of Mechanical & Industrial Engineering, University of Toronto, Toronto, ON, Canada. Before joining the University of Toronto, she was a Postdoctoral Associate with the Massachusetts Institute of Technology, Cambridge. Her research interests are centered on understanding and improving human behavior and performance in multi-task and complex situations, using a wide range of analytical techniques.

M. L. Cummings received the B.S. degree in mathematics from the United States Naval Academy, Annapolis, MD, in 1988, the M.S. in space systems engineering from the Naval Postgraduate School, Monterrey, CA, in 1994, and the Ph.D. degree in systems engineering from the University of Virginia, Charlottesville, in 2004.

A Naval Officer and military pilot from 1988-1999, she was one of the Navy's first female fighter pilots. She is currently an Associate Professor with the Department of Aeronautics & Astronautics, Massachusetts Institute of Technology, Cambridge. Her previous teaching experience includes instructing for the U.S. Navy with the Pennsylvania State University, University Park, and as an Assistant Professor with the Engineering Fundamentals Division, Virginia Polytechnic Institute and State University, Blacksburg. Her research interests include human supervisory control, human-uninhabited vehicle interaction, bounded collaborative human-computer decision making, decision support, information complexity in displays, and the ethical and social impact of technology.