# Developing Performance Metrics for the Supervisory Control of Multiple Robots

Jacob W. Crandall
Dept. of Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, MA
jcrandal@mit.edu

M. L. Cummings
Dept. of Aeronautics and Astronautics
Massachusetts Institute of Technology
Cambridge, MA
missyc@mit.edu

## ABSTRACT

Efforts are underway to make it possible for a single operator to effectively control multiple robots. In these high workload situations, many questions arise including how many robots should be in the team (Fan-out), what level of autonomy should the robots have, and when should this level of autonomy change (i.e., dynamic autonomy). We propose that a set of metric classes should be identified that can adequately answer these questions. Toward this end, we present a potential set of metric classes for human-robot teams consisting of a single human operator and multiple robots. To test the usefulness and appropriateness of this set of metric classes, we conducted a user study with simulated robots. Using the data obtained from this study, we explore the ability of this set of metric classes to answer these questions.

## Categories and Subject Descriptors

J.7 [**Computers in Other Systems**]: Command and Control; H.5.2 [**User Interfaces and Presentation**]: Evaluation/methodology

## General Terms

Measurement, Performance, Human Factors

## Keywords

Multi-robot Teams, Fan-out, Supervisory Control

## 1. INTRODUCTION

Over the last few years, much research has focused on human-robot teams (HRTs) in which a single operator controls or supervises multiple robots. This is a somewhat daunting task as current technologies (in air, ground, and water robotics) require multiple humans to control a single robot. However, it is desirable to invert this ratio in order to (a) reduce costs, (b) extend human capabilities, and (c) improve system efficiency. To achieve this goal, additional research must address a multitude of issues related to both the human operator (i.e., human factors issues), the robots (i.e., artificial intelligence capabilities), and the interactions between them.

One important research agenda is determining the effectiveness of a given HRT in accomplishing a mission. To do so, robust and descriptive metrics must be developed. The first conference on Human-Robot Interaction (HRI 2006) included a paper calling for the development of common metrics for human-robot systems [24]. The authors of this paper argued that metrics should be developed that span the range of missions carried out by HRTs. These metrics should relate to both humans and robots in the team as well as the entire human-robot system (HRS). In this paper, we focus on quantitative metrics for HRTs consisting of a single human operator and multiple robots.

Often, a single metric is sought to evaluate an HRT's effectiveness. However, since metrics of overall system effectiveness vary widely across domains [27] and are typically multi-modal, a common metric for overall system effectiveness is unlikely to be found. However, a *set of metric classes* spanning many aspects (and subparts) of a system is likely to be more generalizable. Loosely, a metric class is the set of metrics that measure the effectiveness of a certain aspect of a system. For example, we might consider the metric class of human performance, which includes metrics of reaction time, decision quality, situation awareness, workload, etc.

We propose that a set of metric classes should have the following three attributes to effectively evaluate HRTs:

1. The set of metric classes should contain metrics that *identify the limits of all agents* (both human operator and robots) in the team.

2. The set of metric classes should have *predictive power*. An HRT might be called upon to perform many different kinds of missions in many different kinds of environments. An HRT that performs well in one environment or mission may not perform well in another environment or mission. Additionally, the teams themselves are likely to change (due to casualty, resource availability, mission needs, etc.). Measuring all such circumstances is costly and, ultimately, impossible. Thus, a set of metrics for HRTs should have some power to predict how changes in environment, mission, and team make-up will affect the team's effectiveness.

3. The set of metric classes should contain *key performance parameters* (KPPs). KPPs are the parameters that indicate the overall effectiveness of the system.


*HRI'07,* March 10–12, 2007, Arlington, Virginia, USA.

Finding a set of metric classes with these three attributes is important for a number of reasons. First, a set of metrics having these attributes can determine the capabilities of a system performing a given mission. In the context of an HRT consisting of a single human operator and multiple robots, such a set of metric classes addresses the question of whether a particular HRT is capable of completing a mission in a satisfactory manner or whether the team's configuration should change. Second, a set of metrics having these three attributes can help determine the levels of autonomy that the robots in the team should employ. This relates to a third reason, which is that such a set of metrics could be used to facilitate dynamic autonomy to a higher degree of fidelity. Fourth, such a set of metrics should identify how changes in system design will impact the system's overall effectiveness. This would both reduce the cost of creating robust HRTs while speeding up their development.

Identifying a set of metrics with these capabilities is a tall order. Nevertheless, we describe initial attempts to do so in this paper. We take the approach of decomposing an HRT into subparts. Measures can be obtained for each of these subparts. Estimates of overall team effectiveness can then potentially be constructed from these measures, even (ideally) when some aspects of the system, environment, or mission change. We demonstrate the potential ability of this set of metric classes via a user study.

The remainder of this paper proceeds as follows. In Section 2, we outline related work. In Section 3, we decompose a single-human multi-robot team into subparts and define metrics for the various subparts. In Section 4, we describe a user study designed to analyze the set of metric classes proposed in Section 3. We present and discuss the results of the user study in Section 5. We offer a concluding remarks and suggest future work in Section 6.

While HRTs of the future will include heterogeneous sets of robots, we focus in this paper only on the homogeneous case. However, the theories developed in this paper appertain to heterogeneous robot teams as well, though additional issues will need to be considered for those teams.

# 2. RELATED LITERATURE

The work of this paper relates to many topics in the literature. We focus on supervisory control of multiple robots, Fan-out, human-robot metrics, and dynamic autonomy.

## 2.1 Supervisory Control of Multiple Robots

In supervisory control [21], a human interacts with automation as the automation acts in the world (see Figure 1). When a human supervises multiple robots, care must be taken to ensure that the operator has the capacity to give adequate attention to each robot or group of robots. Adherence to multiple principles are required to make this possible, including offloading low-level control of the robots to the automation [4, 20, 6, 17], ensuring that the automation is reliable [7], and providing effective user interfaces (see [14, 23]). Predictive metrics are necessary to evaluate these technologies in a cost effective manner.

When a human controls multiple robots, the human must necessarily allocate his/her attention between the various robots or groups of robots. This is related to the concept of time-sharing (see [27, 1]). Metrics from the *attention allocation efficiency* (AAE) metric class discussed in Section 3.2 can be used to assess time-sharing capabilities.

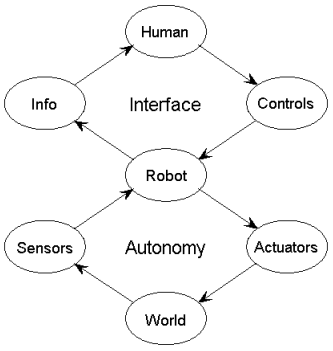


**Figure 1: The two control loops of an HRT consisting of a single human operator and a single robot.**

## 2.2 Fan-out

The term Fan-out (FO) refers to the number of (homogeneous) robots that a single operator can effectively control [16]. One line of research on this topic uses measures of interaction times and neglect times to estimate FO [11, 16, 3]. These metrics have been modified to include the use of wait times [14, 5] and extended (in part) to the domain of heterogeneous robot teams [12]. We analyze how effectively these metrics estimate true FO in Section 5.2.2.

## 2.3 Human-Robot Metrics

Much of the work on metrics for HRTs has focused on the human operator. The most common of these metrics measure situation awareness (SA) (formally defined in [9] and adopted to HRTs in [8]) and operator workload. Various metrics for SA have been devised including SAGAT [9]. Metrics for measuring operator workload include subjective methods (see [27]), secondary task methods, and psychophysiological methods (e.g., [13, 25]). However, metrics for HRTs must go beyond the human operator. Metrics are also needed to evaluate the effectiveness of individual robots in the team as well as the team's overall effectiveness [26].

The work of this paper focuses on combining metrics from various aspects of the HRT to obtain measures of system effectiveness. This is relates to [19], which computes a measure of overall team effectiveness using measures of the individual subtasks performed by the team.

## 2.4 Dynamic Autonomy

Central to the success of an HRT is the level of automation employed by the robots in the team. Sheridan and Verplank's [22] scale of levels of automation has been widely accepted and adapted for use in system design (e.g., [18]). The level of automation can be varied over time (dynamic autonomy) to manage changing operator workload and mission needs (e.g., [17, 2]). Predictive metrics can be used to determine when autonomy levels should be changed.

# 3. A SET OF METRIC CLASSES

We can identify a potentially useful set of metric classes by decomposing an HRT consisting of a single human and multiple robots into subparts. We first decompose a single robot team after which we take on the multi-robot case.

## 3.1 The Single-Robot Case

In the single-robot case, an HRT has the two control loops shown in Figure 1, which is adapted from [3]. These control loops are the control loops of supervisory control [21]. In the upper loop, the human interacts with the robot. The robot sends information about its status and surroundings to the human via the interface. The human synthesizes the information and provides the robot with input via the interface. The lower control-loop depicts the robot's interactions with the world. The robot combines the operator's input with its own sensor data to determine how to act.

The two control loops, though intimately linked, provide a natural decomposition of an HRT of this type. Corresponding to each control loop is a metric class. Metrics that evaluate the effectiveness of human-robot interactions (upper control loop) are in the metric class of *interaction efficiency* ($IE$). Metrics that evaluate the robot's autonomous capabilities (lower control loop) are in the metric class of *neglect efficiency* ($NE$). Note, however, that while these two metric classes are separate, they are in no way independent of each other. A failure in one control loop will often cause a failure in the other control loop.

Many metrics in the literature have membership in the $IE$ and $NE$ metric classes. We focus on a small set of these metrics in this paper.

### 3.1.1 Interaction Efficiency (IE)

Metrics in the $IE$ metric class evaluate the effectiveness of human-robot interactions. That is, they evaluate (a) how well the human can determine the status and needs of the robot, (b) how human inputs affect robot performance, and (c) how much effort these interactions require. One way to estimate $IE$ is by the expected length of a human-robot interaction. This metric is known as *interaction time* ($IT$), which (for the single-robot case) is the amount of time it takes for the operator to (a) orient to the robot's situation, (b) determine the inputs (s)he should give to the robot, and (c) express those inputs via the interface [15]. Related to $IT$ is the metric $WTI$ (wait times during interactions) [14], which is the expected amount of time during interactions that the robot is in a degraded performance state.

Using $IT$ and/or $WTI$ to capture $IE$ infers that shorter interactions are more efficient than longer ones. Since this is not always the case, we might also want to consider metrics that more explicitly measure the performance benefits of an interaction. These benefits can be determined by observing how the robot's performance changes during human-robot interactions, which can be calculated from the mathematical structure *interaction impact* ($II$). $II$ is the random process that describes the robot's performance during interactions [3][1]. It is a function of (among other things) the amount of time $t$ since the operator began interacting with the robot. One metric we can derived from $II$ is the robot's average performance during interactions, which is given by

$$\bar{II} = \frac{1}{IT} \int_0^{IT} E[II(t)]dt, \tag{1}$$

where $E[II(t)]$ denotes the robot's expected instantaneous performance at time $t$ ($t = 0$ is when the interaction began).

[1]For descriptive purposes, we have modified the names of some of the terms discussed in this paper.

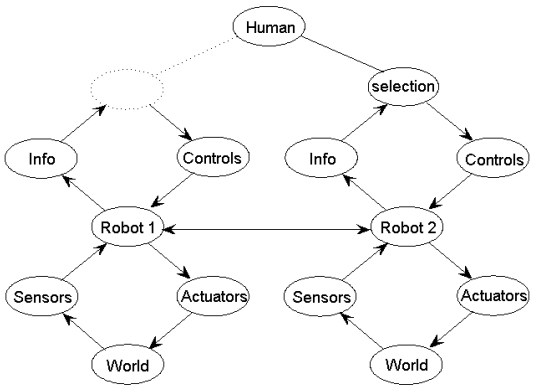


**Figure 2: In multi-robot teams, human attention must be distributed between the robots.**

### 3.1.2 Neglect Efficiency (NE)

The $NE$ metric class consists of metrics that evaluate a robot's ability to act when the human's attention is turned elsewhere. *Neglect time* ($NT$), which is the average amount of time a robot can be ignored before its expected performance falls below a certain threshold [11], is a member of this metric class. One difficulty with this metric is determining the proper performance threshold. Methods for determining the threshold are given in [16, 3]. Like $IT$ and $WTI$, $NT$ does not completely account for the robot's performance. This additional information can be obtained from the mathematical structure *neglect impact* ($NI$), which is the random process that describes a single robot's performance when it is ignored by the operator [3]. From $NI$, we can calculate average robot performance during the time it can be safely neglected using

$$\bar{NI} = \frac{1}{NT} \int_0^{NT} E[NI(t)]dt, \tag{2}$$

where $E[NI(t)]$ denotes the robot's expected instantaneous performance after it has been neglected for time $t$.

## 3.2 The Multi-Robot Case

When a human interacts with multiple robots, the nature of interactions between the operator and each robot in the team remains relatively unchanged except for the important exception depicted in Figure 2. The figure shows a separate set of control loops for each robot. However, unlike the single-robot case, the upper loops are not always closed. To close one of the upper loops, the human must attend to the corresponding robot and neglect the others[2]. Thus, the efficiency with which human attention is allocated among the robots is critical to the team's success. Metrics that capture this notion of efficiency have membership in the *attention allocation efficiency* ($AAE$) metric class.

### 3.2.1 Attention Allocation Efficiency (AAE)

$AAE$ can be measured in various ways including (a) the time required to decide which robot the operator should service after (s)he has completed an interaction with another

[2]We assume that a human *sequentially* attends to the needs of each robot.

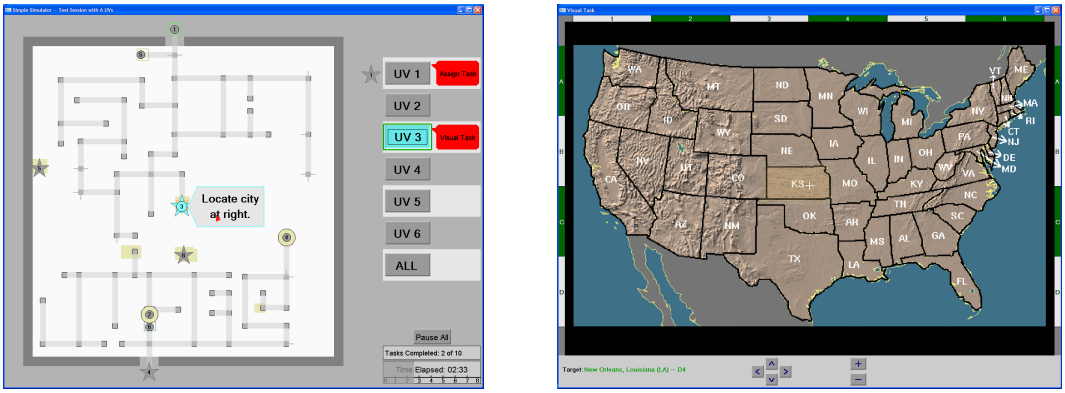


**Figure 3: The two displays used in the experiment. Each was displayed on a separate monitor.**

robot, and (b) the quality of that decision. The former metric is referred to as *switch times* (*ST*s) and has sometimes been considered part of *IT* [16]. We follow this lead in this paper, though it is of itself an individual metric of *AAE*.

Ideally, a metric evaluating the quality of servicing selections made by the HRT would compare the team's actual decisions with what would have been the "optimal" decisions. However, such a measure is often difficult to obtain given the complexity of the situations encountered by HRTs. One alternative metric is to compute the number of *wait times* (i.e., time in which a robot is in a degraded performance state) *caused by lack of operator SA* (called *WTSA*) [14]. In general, teams with higher *WTSA* have lower *AAE*. However, *WTSA* can also be difficult to measure since they must be distinguished from a third kind of wait time, called *wait times in the queue* (*WTQ*) [14]. *WTQ* occur when the human operator knows that a robot is in a degraded performance state, but does not attend to that robot because (s)he must attend to other robots or tasks. The metric *WTQ* is not exclusively from *IE*, *NE*, or *AAE*, though it is affected by all three system attributes.

Figure 2 also shows a connecting link between robots in the team. This link captures the notion that robots can communicate with each other. The quality of information passed over these links will in turn affect measures of *IE*, *NE*, and *AAE*. This could possibly define a fourth metric class, though we do not consider it in this paper.

# 4. USER STUDY

To evaluate how effectively sets of metrics drawn from *IE*, *NE*, and *AAE* identify the limits of the agents in the team, predict system characteristics, and provide KPPs, we conducted a user study. In this section, we describe the software test-bed used in the study, the experimental procedure, and the demographics of the participants.

## 4.1 Software Test-bed

We describe three aspects of the software test-bed: mission, interface, and robot behaviors.

### *4.1.1 Mission*

Across many mission types, an HRT operator assists in performing a set of common tasks including mission planning and re-planning, robot path planning and re-planning, robot monitoring, sensor analysis and scanning, and target designation. These generic tasks apply to HRTs with many different kinds of robots, including unmanned air vehicles (UAVs), unmanned ground vehicles (UGVs), and unmanned underwater vehicles (UUVs). We give two time-critical examples: one with UAVs and the other with UGVs.

A human-UAV team might be assigned various intelligence gathering tasks over a city during the night. The team's mission is to perform as many intelligence gathering tasks before daylight as possible. The operator must assist in assigning the various UAVs to the various intelligence gathering tasks. Once the UAVs are assigned tasks, the UAV operator must assist the UAVs in arriving at the (possibly unknown) locations where these tasks are to be performed. This requires the operator to assist in path planning and the monitoring of UAV progress. As more information becomes available about the various tasks, the intelligence gathering tasks must be reassigned and routes re-planned. Once a UAV arrives at the location where the intelligence must be gathered, the operator must scan the UAV's imagery to identify objects of interest.

A human-UGV team might be tasked with a search and rescue mission in a damaged building. The mission goal would be to remove important objects (such as people) from the building in a timely manner (e.g., before the building collapses). To do this, the operator must assign the UGVs to various places in the building and assist them in getting to these locations. As new information about the building and the objects in it become available, the operator must often reassign the UGVs to other tasks. Once a UGV arrives at the location of an object, it would need the operator's assistance to positively identify and secure the object. This could require the operator to view and analyze imagery from the UGVs video feed. After securing the object, the UGV would then need to exit the building to deliver the object.

We sought to capture each of these generic tasks in our software test-bed, which is shown in Figure 3. In our study, the HRT (which consisted of the participant and multiple simulated robots) was assigned the task of removing objects from an initially unknown maze. The goal was to remove as many objects from the area as possible during an 8-minute session while ensuring that all robots were out of the maze when time expired. An object was removed from the building using a three-step process. First, a robot moved to the location of the object (target designation, mission planning, path planning, and robot monitoring). Second, the robot

"picked up" the object (sensor analysis and scanning). As this action might require the operator to perform a visual task (assist in identifying the object in video data), we simulated this task by asking the user to identify a city on a map of United States using *Google Earth*-style software (the graphical user interface is shown in the right of Figure 3). This locate-a-city task was a primary task and not a secondary task. Third, the robot carried the object out of the maze via one of two exits (one at the top of the maze and the other at the bottom of the maze).

The objects were randomly spread through the maze. The HRT could only see the positions of six of the objects initially. In each minute of the session, the locations of two additional objects were shown. Thus, the total number of objects to collect during a session was 22. Each participant was asked to maximize the following objective function:

$$Score = ObjectsCollected - RobotsLost, \quad (3)$$

where *ObjectsCollected* was the number of objects removed from the area during the session and *RobotsLost* was the number of robots remaining in the area when time expired.

#### 4.1.2 Interface

The human-robot interface used in the study was the two-screen display shown in Figure 3. On the left screen, the maze was displayed along with the positions of the robots and (known) objects in the maze. As the maze was initially unknown to the HRT, only the explored portions of the maze were displayed. The right screen was used to locate cities in the United States.

The user could control only one robot at a time. The user designated which robot (s)he wanted to control by clicking a button on the interface corresponding to the desired robot (labeled UV1, UV2, etc.). Once the user selected the robot, (s)he could control the robot by specifying goal destinations and making path modifications. Goal designation was achieved by dragging the goal icon corresponding to the robot in question to the desired location. Once the robot received a goal command it generated and displayed the path it intended to follow. The user could modify this path using the mouse.

To assist the operator in determining which of the robots needed attention, each robot's status was shown next to its button. This status report indicated if the robot had completed its assigned task, found an object, or needed to exit the maze. If no status report was given, the system determined that the robot was progressing adequately on its assigned task.

#### 4.1.3 Robot Behavior

The robot combined a goal seeking (shortest path) behavior with an exploration behavior to find its way toward its user-specified goal. This behavior, though generally effective, was sometimes frustrating to the users as it often led to seemingly undesirable actions (though, as we mentioned, the user could modify the robot's path if desired).

### 4.2 Experimental Procedure

After being trained on all aspects of the system and completing a comprehensive practice session, each user participated in six 8-minute sessions. Teams with two, four, six, and eight robots were tested. In each of the first four sessions, a different number of robots were allocated to the team. In the last two sessions, the conditions (i.e., robot team size) from the first two sessions were repeated. Thus, 18 samples were taken for each robot team size[3]. The conditions of the study were counter-balanced. The participants were paid $10 per hour with the highest scorer also receiving a $100 gift certificate.

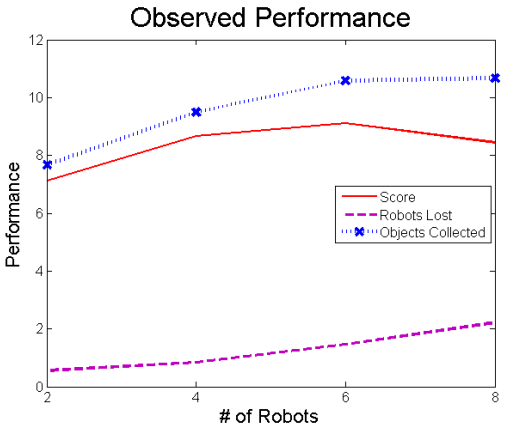


**Figure 4: The mean values of number of objects collected, number of robots lost, and overall score.**

### 4.3 Demographics

Twelve people (one professor, ten students, and one other non-academic person) participated in the study; eight were from the United States, two were Canadian, one was Hispanic, and one was Egyptian. Three of these participants were female and nine were male. The mean age was 27.5 years old with a standard deviation of 8.6 years.

## 5. RESULTS

Data collected from the user study allows us to evaluate sets of metrics (drawn from $IE$, $NE$, and $AAE$) with respect to their ability to identify the limits of the agents in the team, predict system characteristics, and provide KPPs. Before presenting this analysis, we report observations of system effectiveness for each robot team size.

### 5.1 Observed Team Effectiveness

The dependent variables we consider for HRT effectiveness for this user study are those related to Equation 3: the number of objects collected by the HRT over the course of a scenario and the number of robots lost during a scenario. The mean observations for these dependent variables across the number of vehicles is shown in Figure 4.

Figure 4 shows that a 2-robot HRT collected on average just less than eight objects per 8-minute session. The sample mean steadily increases as team size increases up until 6-robots, at which point it appears to plateau. A repeated measures ANOVA revealed marginal significance across robots, $\alpha = 0.05$, $F = (15, 3) = 2.737$, $p = 0.06$. Pairwise comparisons show that 2-robot teams collect significantly less objects than do 4-, 6-, and 8-robot teams ($p \leq 0.001$), and 4-robot teams collect significantly less objects than 6- and 8-robot teams ($p = 0.057$ and $p = 0.035$,

[3]Only 17 samples are available from the 6-robot condition due to technical difficulties.

| | 2 | 4 | 6 | 8 |
|---|---|---|---|---|
| $IT$ | 18.19 | 16.86 | 15.82 | 15.74 |
| $NT$ | 22.26 | 36.63 | 44.67 | 52.22 |
| $WT$ | 8.71 | 26.88 | 45.03 | 67.58 |

**Table 1: Estimated values of $IT$, $NT$, and $WT$ given in seconds per robot team size (the columns).**

respectively). HRTs with six and eight robots are statistically the same.

Figure 4 also shows that the average number of robots lost per session increases as robot team size increases. Robots were lost if they were in the maze when time expired. A clear distinction exists between groupings of 2- and 4-robot teams and 6- and 8-robot teams as demonstrated by a $\chi^2$-test ($\chi^2 = 14.12$, $df = 3$, $p = .033$)[4]. This result is significant as it indicates a performance drop between four and six robots. Thus, while robot teams with six and eight robots collected more objects than smaller robot teams, they also lost more robots.

These results indicate that the HRTs in the user study with the highest performance had, on average, between 4 and 6 robots. Thus, FO for this particular situation appears to be between four and six robots.

## 5.2 Analysis of Sets of Metrics

We now analyze selected sets of metrics drawn from $IE$, $NE$, and $AAE$ with respect to the three attributes listed in the introduction. Namely, we want to determine how well these metrics determine the limits of the agents (both the human and the robots) in the team, predict system characteristics, and provide key performance parameters (KPPs). We analyze each attribute separately.

### 5.2.1 Limits of the Agents

The observed values of $IT$, $NT$, and $WT$ (the average wait time per interaction-neglect cycle) are given in Table 1. We used the following heuristics to calculate them:

- $IT$ was determined by observing clicks on the robot selection buttons as well as other mouse activity. Estimated switch times, which were about 1.7 seconds in each condition, are included in this measure.
- $NT$ was determined to be the time elapsed between the operator's last interaction with the robot and the time at which the operator again interacted with the robot or the robot reached its designated goal location.
- $WT$ was determined to be the average time a robot waited to be serviced after it reached its goal. Thus, both $WTQ$ and $WTSA$ are included in this measure. If a robot did not reach its goal before the operator chose to service it, we assumed that no wait times accrued.

Previous discussions of operator capacity based on the measures $IT$, $NT$, and $WT$ are given in [16, 14, 5]. We provide analysis of operator capacity using these measures for our specific study.

In a 2-robot team, Table 1 shows that, on average, a robot was serviced for about 18 seconds ($IT$), then moved productively toward its goal while being neglected for about 22 seconds ($NT$), and then waited for operator input for a little less than 9 seconds ($WT$). Thus, the robot was either actively pursuing its goal or being serviced more than 82% of the time. This indicates that the operator was usually able to provide adequate attention to both robots. However, as the number of robots in the team increased, the amount of time the operator was able to give adequate attention to each robot decreased noticeably. In 8-robot teams, the user was typically unable to attend to the needs of each robot in the team as each robot spent about half of its time waiting for operator input. As a result, as team size increased, the number of objects collected reached a plateau while the number of robots lost continued to increase (see Figure 4).

We can make observations about the limits of the robots by observations of $NT$. In the 8-robot condition, when interactions with each robot were infrequent, $NT$ was about 53 seconds. Since each robot received little attention from the users in this condition, this value is largely a function of the average time it took for the robots to reach their goals. Thus, it appears that a main limitation of the robots' autonomy was its dependence on user specified goals. Thus, future improvements in robot autonomy could include giving the robots the ability to create their own goals or initiatives.

### 5.2.2 Predictive Power

In this context, predictive power is the ability to determine how the HRT will perform in unobserved conditions. Thus, metrics are predictive if measures obtained in one condition (e.g., a fixed robot team size) can be used to accurately calculate measures for other (unobserved) conditions (e.g., other robot team sizes). Predictive metrics have two attributes. First, they are *accurate*, meaning that their predictions are close to the actual measures we would have observed in that condition. Second, they are *consistent*, meaning that the predictions are accurate regardless of the observed condition(s). These attributes can be assessed in both *relative* and *absolute* ways [27].

In this subsection, we will analyze the ability of three methods to predict FO and system effectiveness as robot team size changes. Each method uses a different set of metrics drawn from the $IE$, $NE$, and $AAE$ metric classes.

**Predicting Fan-out.** The first method, which was presented in [15], estimates FO using:

$$FO = \frac{NT}{IT} + 1. \tag{4}$$

Thus, this method assumes FO is determined by the number of interactions that can occur with other robots while a robot is being neglect.

The second method, presented in [14], adds wait times to Equation (4) so that FO is computed using:

$$FO = \frac{NT}{IT + WT} + 1, \tag{5}$$

where $WT = WTQ + WTSA$.

The third method is a performance-based method described in [3]. This method uses the metrics $IT$, $\bar{I}I$, $\bar{N}I$, and $NT$ (though $IT$ and $NT$ are determined in a slightly different fashion than in Table 1). In short, values of $IT$, $\bar{I}I$, and $\bar{N}I$ are enumerated for all possible values of $NT$. For each possible tuple $(IT, NT)$ a corresponding average robot performance $\bar{V}$ is calculated using

$$\bar{V} = \frac{1}{IT + NT}\left(IT \cdot \bar{I}I + NT \cdot \bar{N}I\right).$$

[4]The $\chi^2$-test for significance was used in this case since the data violated the assumptions on an ANOVA test.

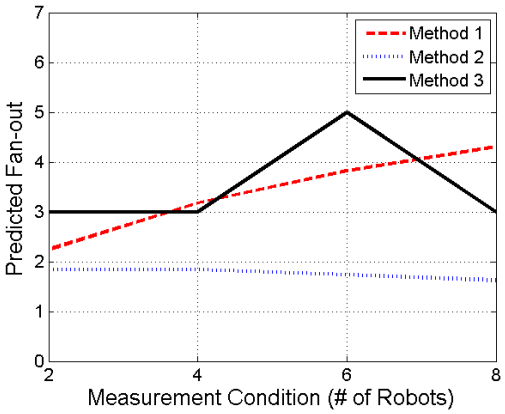


**Figure 5: FO predictions using measures obtained from observing 2-, 4-, 6-, or 8-robot teams.**

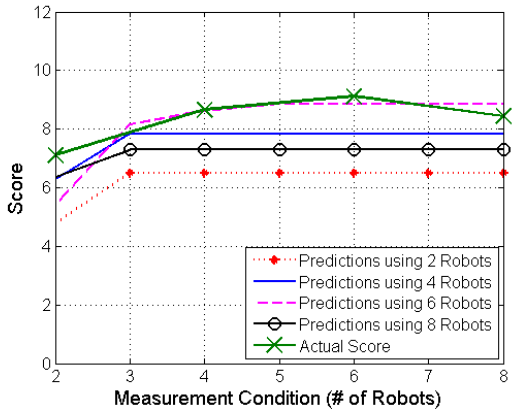


**Figure 6: Predictions of system effectiveness based on metrics obtained using 2-, 4-, 6-, and 8-robot teams.**

Each robot in the team is then assigned its own $(IT, NT)$ tuple such that the sum of robot performances is maximized given the constraint: $NT_j \geq \sum_{i \neq j} IT_i$ for all $j$ (where $NT_i$ and $IT_i$ are the neglect and interaction times assigned to robot $i$). This calculation is made for teams of all sizes. FO is the point where performance peaks or plateaus.

FO predictions for each of these methods (using the values shown in Table 1) are shown in Figure 5. In the figure, the x-axis represents the robot team size that was observed and the y-axis shows the resulting FO prediction. None of the methods consistently predicts the true FO (which, as we discussed previously, was between four and six robots). Method 1 predicts FO to be anywhere from 2.45 (when observing two robots) to 4.32 (when observing eight robots). Thus, this method is not consistent due to variations in the estimate of $NT$. Method 2's FO estimates, though pessimistic, are relatively consistent. This is an interesting result since Method 2 is the only method of the three that uses a metric with (partial) membership in the $AAE$ metric class (other than combining $ST$ with $IT$). It appears that the variabilities in $NT$ are counteracted by $WT$. Future work should investigate whether this trend holds in other contexts. Method 3, also provides a pessimistic estimate, though its predictions are consistent except for the 6-robot team condition (at which point it gives a good estimate of FO). We illustrate why this method fails by analyzing its ability to predict system effectiveness.

**Predicting System Effectiveness.** Methods 1 and 2 use temporally-based methods that only predict the number of robots a team should have. They do not predict what a team's effectiveness will be (for any robot team size). Method 3, however, was designed to predict system effectiveness [3]. These predictions for the HRTs observed in the user study are shown in Figure 6. A set of predictions for each observed robot team size is given. We make several observations.

First, these predictions are not consistent in the absolute sense, though they are in the relative sense. While the predictions follow similar trends they do not always even accurately predict the observed conditions. Second, the figure shows that these predictions are on scale with the actual scores. However, the predictions plateau much sooner than the actual observed scores do. This shortcoming appears to be caused (at least in part) to the method's reliance on average values of $IT$, $NT$ and robot performance. Future work should investigate this claim. Lastly, though not demonstrated here, this method can make vastly incorrect predictions under certain situations [3]. Some reasons for these failures are addressed in [10].

In closing our discussion on predictive power, we make the following observations. First, it appears that predictive tools that use measures from all three metric classes ($IE$, $NI$, and $AAE$) may be better at providing consistent predictions. Second, performance-based measures seem to be more desirable than time-based measures as they (a) appear to give more accurate predictions and (b) can predict more measures.

### 5.2.3 *Key Performance Parameters (KPPs)*

The third desirable element of a set of metric classes is that they contain KPPs. Obviously, more than one KPP can exist. However, in the interest of space we discuss just one KPP for this user study, which was the average time it took for a user to to locate a city on the map (part of $IT$). Several users in the study believed that their performance was driven by how quickly they could perform this primary task. Their claim seems to be somewhat valid as the average time it took a user to find a city on the map was negatively correlated ($r = -717$) with the users' score (from Equation 3). Thus, it appears that an effective way to improve these HRTs' overall effectiveness would be to provide the operator with additional aids in locating the city on the map (or, for a real world example, aids for identify a potential target in video imagery). Such aids could include automated target recognition assistance, etc.

## 6. DISCUSSION AND FUTURE WORK

We have advocated that sets of metric classes for human-robot teams be developed that indicate the limits of the agents in the team, provide predictive power, and contain key performance parameters. We presented a set of metric classes and analyzed it with respect to these three attributes. While sets of metrics drawn from this set of metric classes show limits of the agents in the team and contain KPPs, they fall short in the category of predictive power. Future

sets of metrics drawn from these classes and other metric classes should improve upon these results.

## 7. ACKNOWLEDGMENTS

This research was funded by MIT Lincoln Laboratory.

## 8. REFERENCES

[1] M. H. Ashcraft. *Cognition*. Prentice Hall, third edition, 2002.

[2] J. Brookshire, S. Singh, and R. Simmons. Preliminary results in sliding autonomy for assembly by coordinated teams. In *Proceedings of the International Conference on Robots and Systems*, 2004.

[3] J. W. Crandall, M. A. Goodrich, D. R. O. Jr., and C. W. Nielsen. Validating human-robot systems in multi-tasking environments. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 35(4):438–449, 2005.

[4] M. L. Cummings and S. Guerlain. An interactive decision support tool for real-time in-flight replanning of autonomous vehicles. In *AIAA $3^{rd}$ "Unmanned Unlimited" Technical Conference, Workshop and Exhibit*, 2004.

[5] M. L. Cummings, C. Nehme, and J. W. Crandall. Predicting operator capacity for supervisory control of multiple UAVs. *Innovations in Intelligent UAVs: Theory and Applications, Ed. L. Jain*, 2006. In press.

[6] S. R. Dixon and C. D. Wickens. Unmanned aerial vehicle flight control: False alarms versus misses. In *Proceedings of the $12^{th}$ International Symposium on Aviation Psychology*, 2003.

[7] S. R. Dixon, C. D. Wickens, and D. Chang. Unmanned aerial vehicle flight control: False alarms versus misses. In *Proceedings of the Human Factors and Ergonomics Society 48th Annual Meeting*, 2004.

[8] J. Drury, J. Scholtz, and H. A. Yanco. Awareness in human-robot interactions. In *Proceedings of the IEEE Conference on Systems, Man and Cybernetics*, Washington, DC, 2003.

[9] M. R. Endsley. Design and evaluation for situation awareness enhancement. In *Proceedings of the Human Factors Society 32nd Annual Meeting*, Santa Monica, CA, 1988.

[10] M. A. Goodrich, T. W. McLain, J. W. Crandall, J. Johansen, J. Anderson, and J. Sun. Managing autonomy in robot teams: Observations from four experiments. In *Proceedings of the 2nd Annual Conference on Human-Robot Interaction*, 2007.

[11] M. A. Goodrich and D. R. Olsen. Seven principles of efficient human robot interaction. In *Proceedings of IEEE International Conference on Systems, Man, and Cybernetics*, pages 3943–3948, Washington, DC, 2003.

[12] M. A. Goodrich, M. Quigley, and K. Cosenzo. Task switching and multi-robot teams. In *Proceedings of the Third International Multi-Robot Systems Workshop*, 2005.

[13] T. C. Hankins and G. F. Wilson. A comparison of heart rate, eye activity, eeg and subjective measures of pilot mental workload during flight. *Aviation, Space and Environmental Medicine*, 69(4):360–367, 1998.

[14] P. J. Mitchell, M. L. Cummings, and T. B. Sheridan. Mitigation of human supervisory control wait times through automation strategies. Technical report, Humans and Automation Laboratory, Massachusetts Institute of Technology, June 2003.

[15] D. R. Olsen and M. A. Goodrich. Metrics for evaluating human-robot interactions. In *NIST's Performance Metrics for Intelligent Systems Workshop*, Gaithersburg, MA, 2003.

[16] D. R. Olsen and S. B. Wood. Fan-out: Measuring human control of multiple robots. In *Proceedings of the Conference on Human Factors in Computing Systems (CHI)*, 2004.

[17] R. Parasuraman, S. Galster, P. Squire, H. Furukawa, and C. Miller. A flexible delegation-type interfface enhances system performance in human supervison of multiple robots: Empirical studies with roboflag. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 35(4):481–493, 2005.

[18] R. Parasuraman, T. B. Sheridan, and C. D. Wickens. A model of types and levels of human interaction with automation. *IEEE Transactions on Systems, Man, and Cybernetics – Part A: Systems and Humans*, 30(3):286–297, 2000.

[19] G. Rodriguez and C. R. Weisbin. A new method to evaluate human-robot system performance. *Autonomous Robots*, 14(2-3):165–178, 2003.

[20] H. A. Ruff, G. L. Calhoun, M. H. Draper, J. V. Fontejon, and B. J. Guilfoos. Exploring automation issues in supervisory control of multiple uavs. In *Proceedings of the Human Performance, Situation Awareness, and Automation Technology Conference*, pages 218–222, 2004.

[21] T. B. Sheridan. *Telerobotics, Automation, and Human Supervisory Control*. The MIT Press, 1992.

[22] T. B. Sheridan and W. L. Verplank. Human and computer control of undersea teleoperators. Technical report, Man-Machine Laboratory, Massachusetts Institute of Technology, Cambridge, MA, 1978.

[23] P. Squire, G. Trafton, and R. Parasuraman. Human control of multiple unmanned vehicles: effects of interface type on execution and task switching times. In *Proceeding of the 1st Annual Conference on Human-robot Interaction*, pages 26–32, New York, NY, USA, 2006. ACM Press.

[24] A. Steinfeld, T. Fong, D. Kaber, M. Lewis, J. Scholtz, A. Schultz, and M. Goodrich. Common metrics for human-robot interaction. In *Proceedings of the 1st Annual Conference on Human-Robot Interaction*, 2006.

[25] J. A. Veltman and A. W. K. Gaillard. Physiological workload reactions to increasing levels of task difficulty. *Ergonomics*, 41(5):656–669, 1998.

[26] E. D. Visser, R. Parasuraman, A. Freedy, E. Freedy, and G. Weltman. A comprehensive methodology for assessing human-robot team performance for use in training and simulation. In *Proceeding of the Human Factors and Ergonomics Society 50th Annual Meeting*, 2006.

[27] C. Wickens and J. G. Hollands. *Engineering Psychology and Human Performance*. Prentice Hall, Upper Saddle River, NJ, third edition, 2000.