Searching for a Quantitative Proxy for Rover Science Effectiveness

Erin Pudenz, Geb Thomas,
Justin Glasgow
Mechanical and Industrial Engineering
The University of Iowa
Iowa City, IA 52242, USA
epudenz│gthomas│jmglasgo
@engineering.uiowa.edu

Peter Coppin,
David Wettergreen
Carnegie Mellon University
5000 Forbes Ave
Pittsburgh, PA 15213, USA
coppin@cmu.edu |
dsw@cs.cmu.edu

Nathalie Cabrol
NASA Ames Research
Center/SETI Institute
Moffett Field, CA 94035-100, USA
ncabrol@mail.arc.nasa.gov

ABSTRACT

During two weeks of study in September and October of 2004, a science team directed a rover and explored the arid Atacama Desert in Chile. The objective of the mission was to search for life. Over the course of the mission the team gained experience with the rover and the rover became more reliable and autonomous. As a result, the rover/operator system became more effective. Several factors likely contributed to the improvement in science effectiveness including increased experience, more effective search strategies, different science team composition, different science site locations, changes in rover operational capabilities, and changes in the operation interface. However, it is difficult to quantify this effectiveness because science is a largely creative and unstructured task. This study considers techniques that quantify science team performance leading to an understanding of which features of the human-rover system are most effective and which features need further development. Continuous observation of the scientists throughout the mission led to coded transcripts enumerating each scientific statement. This study considers whether six variables correlate with scientific effectiveness. Several of these variables are metrics and ratios related to the daily rover plan, the time spent programming the rover, the number of scientific statements made and the data returned. The results indicate that the scientists created more complex rover plans without increasing the time to create the plans. The total number of scientific statements was approximately equal (2187 versus 2415) for each week. There was a 50% reduction in bytes of returned data between the two weeks resulting in an increase in scientific statements per byte of returned data ratio. Of the original six, the most successful proxies for science effectiveness were the time to program each rover task and the number of scientific statements related to data delivered by the rover. Although both these measures have face validity and were consistent with the results of this experiment, their ultimate empirical utility must be measured further.

HRI'06, March 2–4, 2006, Salt Lake City, Utah, USA.
Copyright 2006 ACM 1-59593-294-1/06/0003...$5.00.

Categories and Subject Descriptors

H.5.2 **[Information Interfaces and Presentation]:** User Interfaces – *Ergonomics, Evaluation/methodology, graphical user interfaces (GUI), training, help and documentation, user-centered design.*

General Terms

Measurement, Documentation, Design, Human Factors

Keywords

Human Robot Interaction (HRI), Mobile Robots, Supervisory Control, Teleoperation Interface, Remote Rover Exploration

1. INTRODUCTION

1.1 Measuring Vicarious Science Effectiveness

Using rovers to explore other planets is no longer a novel idea. The Soviet Union landed Lunokhod 1 and 2 to explore the lunar surface in the 1970s [3]. In 1997 NASA landed Sojourner Truth, which navigated on the surface of Mars [8]. More recently, NASA landed two exploration rovers on Mars [10]. As of September 2005 NASA has been successfully operating the Spirit and Opportunity rovers for a combined total of 1200 mission days.

Rovers have also been used to investigate dangerous sites. A rover entered the Three Mile Island power plant to reduce the risk of radiation to workers [16]. The Pioneer rover was designed to enter the damaged Chernobyl reactor [11]. The Dante II rover descended Mt. Spur Alaska to collect gas samples [2].

These missions, while different, are all examples of vicarious science. Vicarious science is simply the process by which operators and scientists remotely explore and assess environments. An increasing number of successful missions suggest that the HRI and rover design community can focus on developing methods to make scientific exploration more effective. Specifically, how can the design of the rover, rover data interface, and training protocols help scientists quickly and accurately form useful scientific conclusions about the remote environment? An important step in addressing this question is to understand how to measure the effectiveness of a particular mission.

There is a call for measurements of scientific effectiveness in all areas of scientific space exploration. This demand is currently being addressed in areas of space telescope and automated spacecraft missions [9]. However, the effectiveness measurements for rover operation has yet to be developed. Previous efforts focused on what scientists perceive via the rover versus what they perceive in the actual environment [12, 13]. This approach allows the development team to identify differences between the information provided by the rover and the information a human would gather at the same location. Rover developers are able to use this knowledge to create a new or altered design. However, this process is not possible in situations where it is too dangerous or remote for humans. Therefore, alternative methods that determine science effectiveness solely by observing science operation in the control room are necessary.

Various rover interface designers have suggested that mission success depends on effective communication between the science team and the rover [1, 4, 6]. Much of this communication emphasizes the rover's motion and navigational tasks. A previous study of scientist behavior during a mission suggests that scientists are actively engaged in debating the meaning of the data returned and actions the rover should take to most effectively achieve their goals [5]. Consequently, the communication among the scientists and between the scientists and the rover may provide important cues to mission effectiveness.

1.2 Mission Objectives

The main function of the scientists is to address the science mission goals. For example, in the study described here there are two primary science-related mission goals. The first is to seek and characterize biota surviving in the Atacama and analyze microhabitats. The second is to determine the physical and environmental conditions associated with identified past and current habitats, including the search for structural fossils, the monitoring of current biological oases and micro-organic communities, and learn how these organisms have contributed to the modification of their environment.

To achieve these goals, the scientists must rely on their individual expertise and experiences with rover control to help them analyze and acquire satellite and rover based data that address the mission objectives. Data analysis consists of searching the data for patterns, determining whether these patterns are consistent with their understanding of the physical environment, forming hypotheses, and summarizing their findings to their peers. Data acquisition consists of analyzing available data to determine the rover's position, deciding where to direct the rover, and deciding what data to collect to test hypotheses or learn about new environments. Individual scientists accomplish much of the data analysis and some of the data acquisition while working quietly at their computer. Consequently, many of the process details are invisible. However, many important portions of the process are observable. These include the record of what the scientists look at and do on the computer, the daily scientist meetings, written scientific summaries, ad hoc discussions among the scientists, the formal rover plan sent to the rover, and the data returned by the rover. The purpose of the work described here is to search some of these sources for variables that measure science effectiveness.

An important challenge in understanding science effectiveness is that it is experimentally difficult to measure the absolute utility of the scientists' actions and statements. Evidence that is compelling to one scientist may be less compelling to another. Quantitative field analysis that disproves a hypothesis made by a scientist does not mean that the scientist should not have formed and supported the hypothesis before the field examination was made. Some conclusions that can be made with certainty may seem predictable, uninteresting, and qualitatively less valuable than other speculative conclusions which seem inspiring and powerful. Ultimately, measuring the importance of a scientific statement is as elusive as measuring the importance of truth.

Nevertheless, it is clearly possible for scientists to be ineffective. They can choose to collect redundant data or overlook the limitations of an instrument and collect data that is useless for making any decisions. Alternatively, they can design clever experiments that provide unique data that succinctly answers an interesting question and leads to new discoveries. They may use the rover's natural strengths to augment their search by filtering data before it is returned, saving precious bandwidth. More often, science activity is neither ideal nor worthless but effective to some intermediate degree. If that elusive parameter could be measured, it would be very useful to both the scientists and rover designers.

Without effectiveness measures, participants in and observers of vicarious science can describe but not evaluate. After a mission of a rover prototype it may be clear that lessons were learned, but how much progress had been made towards an effective mission since the last mission? Did the investment in the mission pay a handsome return or is the design and science team refined beyond the point of diminishing return for further missions? Would more design iterations and practice be advantageous? Science effectiveness does not simply depend on the scientists or the rover, but is a product of the complex interaction between both parts of this interconnected system. A metric of science effectiveness would allow researchers to isolate suboptimal interactions between and within the human and machine components by focusing attention on particular phases of the mission, particular instruments, and particular behaviors.

The vagaries of the scientific creative process limit the ability to establish a precise metric for science effectiveness, but it is possible to define performance metrics that roughly track scientific productivity. If the interaction between scientists and rover shows increased productivity, then there is probably an increase in science effectiveness. This allows measurement of quantitative parameters that indirectly correlate to science effectiveness instead of directly measuring science quality.

This study focused on measuring different parameters under conditions that normally produce improvement in science effectiveness. If the examined parameters change in the predicted direction, this supports their correlation to science effectiveness. The nature of this mission introduces a number of confounds meaning any potential parameters must be experimentally verified. Finding reasonable quantitative proxies for science effectiveness is an important goal and this research is the next logical step toward attaining that goal.

1.3 Effectiveness Proxies

This experiment explored six proxies of science effectiveness. The first concerned the interactions between the scientists and the rover programming interface. EventScope was the software used by the scientists to formally record the rover plans. A rover plan was a list of tasks that the scientists directed the rover to perform. The plan was formally constructed by programming the list of tasks into EventScope which created an XML plan. A task was defined as a group of rover actions that, when performed together, achieved a single operation. A task was such things as taking a series of pictures, driving to a new location, plowing into the ground. The hypothesis was that as the science team became more effective, it would take less time for the team to enter the plan.

The second proxy was the rover plan itself. The hypothesis was that as the science team became more complex in their planning, the resulting rover plan would become more complex.

The next proxy addressed the potential problem that more complex programs would require more time to create. However, at the same time the science team should become more efficient at creating the XML code. Therefore, it was necessary to measure the ratio between the time needed to program the rover and the number of tasks in the rover plan. The hypothesis here was that the ratio would decrease with improved science effectiveness.

The fourth proxy was based on the premise that when the science effectiveness was high, the scientists would have more scientifically interesting things to talk about. Consequently, they would talk more about the data from the rover. The hypothesis was that the percentage of scientific statements about rover data would increase with effectiveness. Other statements, such as those about satellite data, are critical to a successful mission; however these should be a lower priority to the team if they are receiving new and interesting rover based data.

The fifth proxy was that as the scientists became more refined in their technique for acquiring data, the data returned would become more productive. Consequently the number of comments made about each byte returned from the rover should increase.

The sixth proxy related to the relative effectiveness of each type of data acquisition instrument on the rover. As the rover/scientist system evolves, a natural balance will be struck between the size of the data package allocated to each instrument with the value of the information it provides, within the limits of the system. As the science operations become more effective, less useful data will be requested less frequently and interesting data will be requested more frequently. The hypothesis was that the ratio of scientific statements per byte of returned data for each instrument will gradually evolve to become equal across instruments.

The observed rover development test consisted of two separate weeks of operation during year two of a three-year rover development process. Between the first and second week, the rover improved, the operator interface changed, and the science team gained experience. All these factors suggest improved science effectiveness during the second week of operations. The purpose of this study was to determine if one or more of the science proxies improved between the two weeks of the mission.

2. METHOD

2.1 Background

The experiment was conducted as part of the Life in the Atacama mission of 2004, in which the rover, Zoë, explored two distinct arid regions in the Chilean Atacama Desert. The science team used the rover to study each distinct region for seven consecutive sols. In this experiment, each sol represented one operation day. The goals of the mission were to look for signs of life in the desert and to understand the desert habitats.

The science team was made up of six scientists the first week and eight scientists during the second. Each team was composed of scientists from different disciplines. The first team consisted of three scientists specialized in biology, two in geology, and one in spectroscopy. The second team consisted of three scientists specialized in biology, four in geology, and one in spectroscopy. The number of scientists available for any sol varied, but was typically five during the first week and seven during the second week. The team, with the exception of one biologist during week two, worked from a mission control room set up in Pittsburgh, PA where they analyzed the data collected by Zoë and created each sol's rover plans. The biologist who was not present in the control room during week two communicated with the other scientists via emails and conference calls.

As part of a simulation of a planetary mission, the team received satellite images of the area near the simulated site (an elliptical area with a long axis of approximately 15 km) before they received the first data from the rover [8]. From this orbital data, the scientist-created data was generated. The scientist-created data was based off satellite maps and indicated possible areas to visit and the general geological structure of the environment. Upon "landing" at each site, the rover returned a high-resolution panoramic image of its surroundings. The scientists could view this data and all available data on a password-protected website. The scientists analyzed the available data and had until the next sol to create and upload a new rover plan. The rover executed the plan and, in the evening, sent newly acquired data to the scientists. The cycle then repeated for the duration of the week.

At 3:00 p.m., the scientists met as a group to finish summarizing data from the previous sol and began drafting a rough outline plan for the next sol. After completing the rough outline, the scientists waited for the new data, which typically arrived around 7:00 p.m. The scientists briefly examined the new data and then met to review the data and make the plan for the next sol's operations more concrete. Then, the team divided into small groups and individuals to analyze the new data. Around 10:00 p.m. the scientists met to finalize the plan and to enter it into EventScope. This programming in EventScope is completed by the lead science team member. The scientists normally finished the plan between 1:00 a.m. and 3:00 a.m. They then break until around 11:00 a.m. They spend until 3:00 p.m. summarizing their finding from the previous day. At 3:00 p.m., they met and started the process over.

The scientists used EventScope to create and upload the rover plan. After determining the area of interest, the team would specify what data they wanted Zoë to collect, along with subtasks required to perform the task. These subtasks included such items as determining location, camera position, and filter type.

During both weeks of the mission, cameras mounted in the ceilings, microphones on the tables and lapel microphones worn by the scientists recorded all the activities and conversations that occurred in the science operations room. Every five minutes, the experimenters manually recorded each scientist's location, activity task, and membership in any conversational groups.

Data collection for the mission also included collecting the webserver log, EventScope log, and rover plans. The webserver log recorded all data uploads to and downloads from the password protected mission website. The EventScope log recorded the activities performed in EventScope. The rover plans are the tasks that were given to Zoë to perform requested operations.

2.2 Webserver Log

The webserver log included each data request received by the server, the time of the request, the file requested, and the IP address of the requester. This data was loaded into a database, where administrative web hits were removed and different copies of the same data were consolidated. Each file was associated with its generating instrument. This database provided the final count of how much data of each type was available to the scientists.

2.3 EventScope Log

Analysis of the EventScope log indicated the amount of time the scientists used EventScope. This log recorded the date, time, and source of each operation, as well as, the tool used and the operation itself. The log was analyzed by dividing it into activity clusters, blocks of time spent working in EventScope. Each cluster was defined as a period of activity with no breaks between actions lasting longer than 5 minutes. The total amount of time on EventScope for each sol was determined by summing the duration of all activity clusters.

2.4 Rover Plan

The rover plans were analyzed by counting the number of tasks requested for each sol. Written in XML, these codes included the initial state of Zoë, the desired tasks to be performed, and the final location of Zoë. Each task includes all subtasks that are needed to obtain the operation. An example of a task would be to collect a panorama; this may involve taking 218 images. Although, it requires a fairly large amount of time, it is considered one task because it returned one complete piece of data. Another task may be to collect a workspace. This would only require one picture to be taken. Thus, the amount of code required to perform each task was not considered in the analysis.

2.5 Time-to-Task Ratio

The time-to-task ratio was the duration of the activity clusters divided by the number of rover tasks for each sol. No time-to-task ratios are available for Sols 7 or 14, because these were the last sols of each mission week and no new plan was generated.

2.6 Total Statements

After the mission was completed, the 5 minute log that recorded the scientists' activities was used to determine peak times of scientific discussion. From these times, the audio tapes were pulled based off of the scientific density of conversations and transcribed. Sixty-one tapes were transcribed into 22,492 lines of transcript. Six students analyzed the transcripts and enumerated each statement made by a scientist that referred to, described, or generalized the mission data. Other data recorded was the date, time, scientists, and task at hand. Each student was supervised and the quality of the final product was comparable to each other.

2.7 Statements per Byte

The number of bytes per each instrument was counted in [7]. There were 1,007MB (559MB without stereo redundancy) returned in the first week and 515MB (266MB without stereo redundancy) returned in the second week. This tally includes all the data transmitted from the rover instruments, but does not include the satellite data nor the scientist-created data, since these data sets were not included in the bandwidth constraints that were imposed on the science team. The data includes both a count of the bytes with a redundancy due to stereo collection and without the redundancy of stereo collection. The bytes per data type were calculated to better understand the cost of each type of data regarding the mission's limited bandwidth. Statements per byte were also analyzed. The statement per bytes were calculated based on the number of statements pertaining to the data types and the number of bytes returned from the rover with the redundancy of stereo since this would be included in the bandwidth constraints.

2.8 Statements per Data Type

There are several different types of data collecting instruments on Zoë. The fluorescence imager (FI) uses different types of dye to detect microbiology under the rover. The stereo panoramic imager (SPI) camera returns panoramic images around the rover. The workspace (WKSP) camera returns an image providing transitional context between the FI and the SPI images. The spectrometer was attached to collect spectral information. (Refer to [14] and [15] for a more detailed account of the data collecting instruments on the rover.) The scientists also received local weather data and satellite images from the general location of the rover. In some cases, the scientists would use these data to annotate satellite images or to create other secondary data products. The number of statements per data type was calculated for each of these data types.

3. RESULTS

3.1 EventScope Log

Table 1 displays the time spent using EventScope to develop the rover plan each week. A paired t-test indicates that there was no significant difference in the amount of time the scientists spent programming in each week (t(5) = 0.776, p = 0.4731).

Table 1. Data Analysis for Each Sol

Sol Number	Total Time using EventScope	Number of Tasks in Rover Plan	Time-to-Task Ratio
1	2:08:19	25.00	0:05:08
2	3:35:18	22.00	0:09:47
3	3:03:56	26.00	0:07:04
4	3:29:42	30.00	0:06:59
5	4:08:46	19.00	0:13:06
6	2:24:00	21.00	0:06:51
7[a]	1:35:14	0.00	N/A
Summation	18:50:01	143.00	N/A
8	3:51:11	104.00	0:02:13
9	1:19:58	129.00	0:00:37
10	2:29:53	64.00	0:02:21
11	4:25:44	94.00	0:02:50
12	2:10:59	142.00	0:00:55
13	1:33:27	25.00	0:03:44
14[a]	0:08:44	0.00	N/A
Summation	15:51:12	558.00	N/A

[a] Not used in calculations because no rover plan was created.

3.2 Rover Plan

Table 1 also provides the number of individual tasks in the rover plan each week. The scientist-generated sequences contained more tasks in the second week than in the first week (paired-t(5) = -3.85, p = 0.01).

3.3 Time-to-Task Ratio

As both Table 1 and Fig. 1 suggest, the time-to-task ratio was significantly higher for week 1 than week 2 (paired-t(5) = 3.93, p = 0.01).

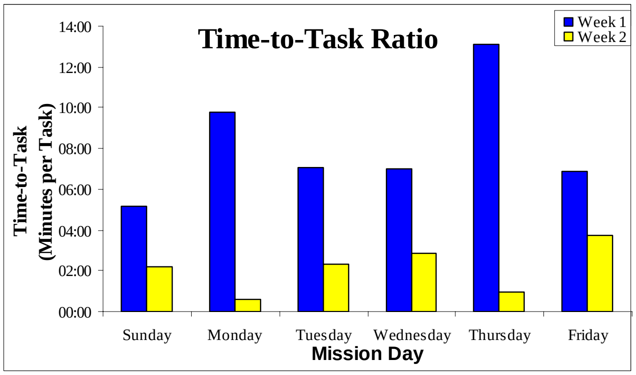

Figure 1. The Time-to-Task Ratio for each mission day in week 1 and week 2. Time-to-Task Ratio was calculated by taking the time spent in EventScope divided by the number of tasks in the rover plan that was entered into EventScope.

3.4 Total Statements

There were 4,963 statements recorded for the two weeks. From that, 361 statements were removed from the count because they were made by individuals outside the science team, such as members of the EventScope support team. After this correction there was 4,602 total statements made by the science team; 2,187 from week 1 and 2,415 from week 2.

3.5 Statements per Byte

In week 2, 51 percent (48 percent without redundancy) less data was returned than in week 1. Table 2 breaks the data returned into instrument categories.

Table 2. Megabytes for Each Data Type

Data Type	Week 1 (Megabytes)	Week 2 (Megabytes)
FI	269	112
SPI	224	124
Weather	7	1
WKSP	27	24
Spectrometer	2	1

Figure 2 and Fig. 3, illustrate how the statement per byte for different types of data compared between week 1 and week 2. Figure 3 shows the same information as Fig. 2 without the spectrometry data. In all cases, there was a higher statement per byte in week 2 than in week 1.

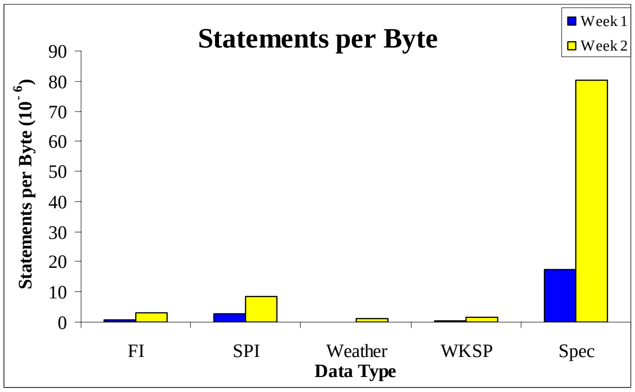

Figure 2. The distribution of the number of scientific statements made regarding each byte of rover data returned by data type.

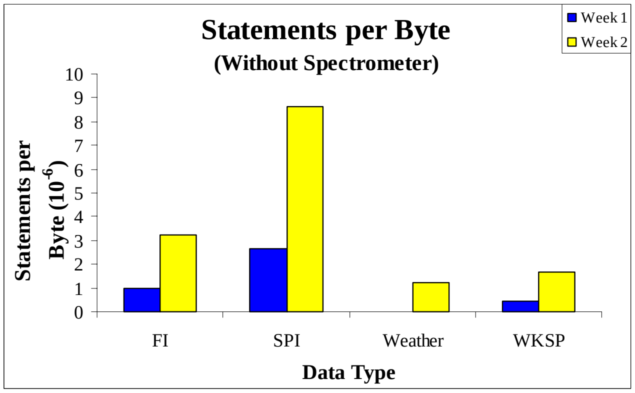

Figure 3. The distribution of the number of scientific statements made regarding each byte of rover data returned by data type sans spectrometer.

3.6 Statements per Data Type

There were 5,849 statements regarding the different data types available (2,811 in the first week and 3,038 in the second week). This number is more than the actual numbers of statements, because some statements referred to more than one type of data at a time. Figure 4 shows the relationship between the statements said about each type from the first week of mission to the second week.

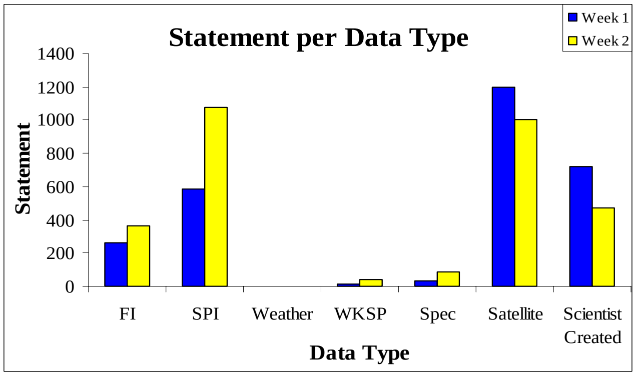

Figure 4. Scientific statements made regarding different data types.

Figure 5 and 6 show the percentage of statements spent on each data type in week 1 and in week 2. Sixty-eight percent of the total statements about data reflected orbital data compared with 48 percent during the second week. Orbital data included satellite data and the scientist created data since most scientist-created data were annotated orbital images.

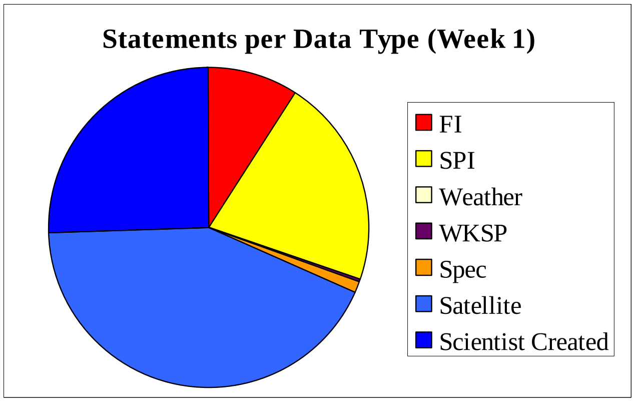

Figure 5. Percent of statement for each data type during week 1. Orbital data includes satellite images and scientist-created data types.

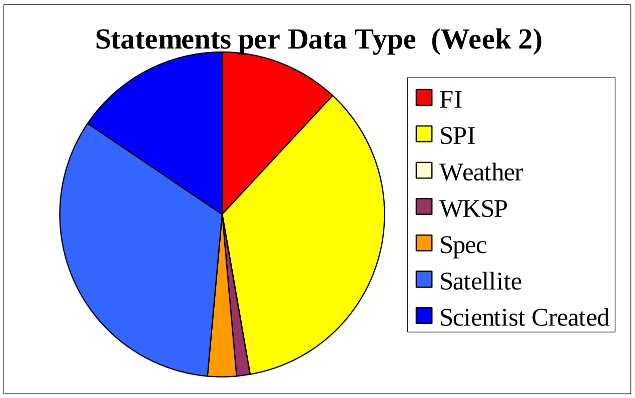

Figure 6. Percent of statement for each data type during week 2. Orbital data includes satellite images and scientist-created data types.

4. DISCUSSION

The results indicate that when operator/rover system effectiveness improved between weeks 1 and 2 the proxies responded as hypothesized. Although there was no significant change in the amount of time spent programming the rover, the number of tasks created in the rover plan was significantly higher in the second week than in first week of the mission. The time-to-task ratio was substantially lower for the second week by approximately a factor of four. The number of statements was nearly consistent from week 1 to week 2, but the number of bytes returned by the rover dropped approximately by a factor of 2. Consequently, the number of statements per byte increased from week 1 to week 2 by approximately a factor of 2. Although the statements per data type did not tend towards uniformity from week 1 to week 2, there was a clear shift in the scientists' attention away from the satellite information towards the rover information.

4.1 Implications Rover Plan Analysis

The increase in the number of tasks in each rover program may be at least partly due to changes in the EventScope interface during the second week of the mission. An added template feature allowed the team to store and recall past task sequences. By using the templates, the science team started the programming sequence with a working plan and could increase the plan complexity as desired. This allowed the team to create more complex programs in less time. The other feature was the pinning operation, which allowed the scientists to designate a data collection point by placing a virtual pin on a digital elevation map textured with satellite images. This eliminated the need to manually determine the coordinates in order to create the task.

Even with the improved effectiveness of the programming interface, there was no consequential reduction in time-spent programming. This suggests that the limiting factor on plan complexity was what the scientists could program in the available time. The simple fact that the team sacrificed rest to stay past 2:00 a.m. to complete the rover plan supports this assertion. Clearly, the team preferred data collected by the more complex programs, which justified to them working the long hours.

Some evidence indicates that the programs in week 2 were not more complex. Each evening, the science team prepared a written support document to send to the field team describing the rover plan created that day. The number of statements in these support documents was not longer in week 2 than in week 1. The explanation for this phenomenon is that the scientists could often represent an entire new task in the rover code with a single modifier in the support document. For example setting five new color wheels may be coded as five tasks in the rover plan, but is one task with five modifiers in the support document.

On the whole, the evidence for a more complex rover program is more compelling than the evidence against it. Further analysis is required to determine the program complexity relative to some absolute standard. The time required to generate a rover plan is probably not a good measure of science effectiveness. This parameter is highly sensitive to the tools the scientists use to generate the plan and how the observers define the planning task. Lastly, as observed in this experiment, the planning task is easily limited by available time. The length of the program seems to correlate with the science effectiveness, but as a proxy, this metric suffers from great variability in task definition from rover to rover. Also, measuring the length or complexity of a program may not be directly correlated to scientific efficiency. In some cases it is necessary to create shorter, simpler plans in order to collect a specific piece of data essential to the mission at hand. It may, however, serve as a useful comparative measure. The programming time-to-task has the advantage of face validity, but if programming time remains constant in other experiments, this proxy is just a restatement of the program length.

4.2 Implications Science Statement Analysis

The enumerated scientific statements indicate that the science team generated roughly the same number of statements each week. There are a number of reasons why this might occur. One is that the science team was time-limited. There was a limited amount of "air time" that each scientist could have to express their ideas, because most of the statements occurred during team meetings and the meetings have an informally fixed duration. Thus, the scientific statements filled the available time, which led to the number of statements being roughly equal.

With a 50% decrease in the number of bytes returned from week 1 to week 2 [7] and the steady rate of scientific statements, the statements per byte increased. This is consistent with the idea that the team learned what data was useful and what was not. Consequently, statements per byte increased because the returned data was of higher quality. When the scientists found what was useful, the size of the data set decreased, more closely approaching the target bandwidth. The scientists also spent more time talking about each byte of data in every data type. Although an alternative explanation for the increased statements per byte is that the actual site of week 2 was more scientifically interesting than the site of week 1, the science team never indicated that the sites actually had a different level of scientific interest.

The number of statements per all types of data increased from week 1 to week 2. Some data types experienced a larger increase than did others. In particular, statements per byte involving spectrometer data nearly quadrupled (Fig. 3). A factor to consider is that the spectrometers returned more data during week 1 but not all of this data was of scientific use as the spectrometers were not always operational. There were fewer bytes of data in every category in week 2 than in week 1. A decrease in the number of bytes returned drastically increases statements per byte for that category. Also, some of the data types tend to be more complex to analyze the results than others. Thus, it should be assumed more complex data types needed to be explained more thoroughly. This sensitivity may be decreased by pooling several data categories. One possible combination is to compare the statements per byte delivered from the rover with the statements per byte for the orbital data. The orbital data consists of the satellite data and the scientist-created data. The majority of the scientist-created data was derived from satellite maps. The rover data consists of all the data returned by the rover each sol, including the FI, SPI, WKSP, and spectrometer data. The amount of rover data was constrained by the bandwidth of the transmission along the communication channel to the rover; whereas, the amount of orbital data available was unconstrained.

Examination of the usefulness of collected data based on scientific statements showed that the scientists emphasized rover data more in week 2 than in week 1. Orbital data was referenced in 68% of the week 1 scientific statements but only 48% of week 2 statements. This supports the idea that the data returned by the rover during the second week provided more interesting information than the data from week 1. Orbital data is a rich source of information and the scientists can always find something interesting to discuss about the satellite data. However, the rover data has no guarantee on being interesting, but the more effective the science, the more likely the rover will return interesting data. Only when the rover returned interesting data will the scientists discuss that data rather than discussing the bigger picture provided by orbital data.

It is important to note that although the orbital data is full of useful knowledge, there are limitations to what it can explain about a region. The rover data helps fill in these voids. While the orbital data shows the big picture of the region, the rover data is able to provide a more in depth understanding on a specific area of the larger region. The rover data is essential to complete the goals of this mission of finding microorganisms, identifying microhabitats, and showing how organisms live in the Atacama Desert. The purpose of this study is to measure rover science effectiveness. Consequently, it seems reasonable to expect that the scientists will use information provided by the rover, otherwise the need for the rover is decreased.

In summary, the observations do not support the hypothesis that the absolute number of scientific statements would increase as science effectiveness improved. Instead the number of statements was not significantly different. The hypothesis that the number of statements per byte would increase was supported; however, the cause appears to be the decrease in the number of bytes returned rather than increased scientist efficiency. Rather than arrive at the counter-intuitive conclusion that a useful proxy for science effectiveness is the reduction in data delivered from the rover, we prefer to conclude that the statements per byte is not a useful proxy for science effectiveness. The hypothesis that the statements per byte for each data type would reach a uniform value was not supported. However, this analysis has led to the discovery that the percentage of statements made about data returned from the rover may be a useful measure of science effectiveness, particularly in cases where there are other rich sources of data competing for the scientists' attention.

4.3 Increase in Science Effectiveness

The preceding analysis was based on the premise that science effectiveness improved from week 1 to week 2. Several observations support this assertion. One is the changes added to EventScope during week 2. The added software features, templates and pinning operations, led the team to be more efficient in entering in the rover plan.

The effects seen may also have been due to a change in the operational capabilities of the rover itself. The rover had the same general mechanical features in weeks 1 and 2. In week 2, software improvements led to the rover being capable of more autonomy than in week 1. For the most part, the science team was unaffected by improvements to the rover autonomy. Their task consisted of creating a plan, uploading the created plan, and downloading the data from the rover. The increase in autonomy may have increased the rover's predictability and consistency.

Perhaps the greatest evidence for increases in system effectiveness was the development of the templates. Throughout the two week period, the science team developed and refined a series of standard data packages for different operations. During the first week, the content of these packages varied considerably, but by the end of the second week, the changes were smaller and less frequent. This pattern supports the premise that scientists gradually learned how to effectively use the rover to do science.

4.4 Potential Confounding Factors

An important confounding factor from week 1 to week 2 was the science team size. In week 1, the team consisted of six members and in week 2, it had eight members. This change in size may have caused the increase in science effectiveness. However, as shown in the results, there was approximately the same amount of scientific statements from week 1 to week 2. This may have been due to the limited amount of time the team had to discuss their findings as a group. Although there were more members in week 2, the group as a whole was still limited in total time to perform their analysis. In addition, the different composition of the team helped to eliminate the increased experience curve. Any learning and increased efficiency found is due to an increased learning and/or efficiency as a whole, not from individual contributions.

The sites the team studied through the rover were different from week 1 to week 2. Arguably, one site may have had more scientifically interesting data than the other. There was no indication in the comments made by the scientists that one site was less interesting or that they had exhausted their ability to analyze the site. Both weeks ended with unanswered questions and, aside from the physical exhaustion, it seemed likely that the science teams would have been glad to have further opportunities to study the area with the rover. Interestingly, the second site was that of a drier climate. Thus, it could be argued that any reports of finding life in week 2 would be of a greater achievement than in week 1 because life clusters are typically found around water.

5. CONCLUSION

Of the six candidate proxies for science effectiveness, the most promising appear to be programming time per rover task and the percentage of scientific statements concerning rover data. As science effectiveness increases, programming time per rover task should decrease and a greater percentage of science statements should emphasize data from the rover. The scientist behavior and operator interface will evolve to improve science effectiveness. The scientists will evolve standard approaches to collecting data with a particular rover. As the data collected becomes more effective from a scientific standpoint, the scientists will spend more time discussing it, provided that other rich, competing data sources are also available.

Although these findings are based on one rover in one experiment, the measures are applicable to a wide range of rover applications involving vicarious science. It is possible that similar measures would also be useful for other search-directed applications, such as rovers for search-and-rescue and military scouting. Because these are largely empirical results, however, their ultimate utility must be tested with repeated application on a wide variety of systems.

6. ACKNOWLEDGMENTS

This work is partially supported by NASA's Applied Information Systems Research Program under NASA grant NAGW5-11981 and NNG05GA51G and the Astrobiology Science and Technology for Exploring Planets under grant NAG5-12890.

5. REFERENCES

[1] Backes, P., et al. The Science Activity Planner for the Mars Exploration Rover Mission: FIDO Mission Results. *Proceeding, 2003 IEEE Aerospace Conference, Big Sky, MT.* 2003, 1-15.

[2] Bares, J.E., and Wettergreen, D.S. Lessons from the development and deployment of Dante II, *Proc. 1997 Field and Service Robotics Conference*, Canberra, Australia, 1998.

[3] Burrows, W., E., This New Ocean: The Story of the First Space Age. New York: Random House, 1998.

[4] Fong, T., et al. Operator Interfaces and Network-Based Participation for Dante II. *Proc. SAE Int. Conf. on Environmental Systems*, 1995.

[5] Fong, T. and Thorpe, C. Vehicle Teleopation Interfaces. *Autonomous Robots*, 2001, 11, 9-18.

[6] Fong, T., Thorpe, C., and Baur, C. Advanced Interfaces for Vehicle Teleoperation: Collaborative Control, Sensor Fusion Displays, and Remote Driving Tools. *Autonomous Robots*, 11, 2001, 77-85.

[7] Glasgow, J., Pudenz, E., Thomas, G., Coppin, P., Cabrol, N., and Wettergreen, D. Observations of a Science Team During an Advanced Planetary Rover Prototype Mission, *14th IEEE International Workshop of Robot and Human Interaction Communication*, Nashville, TN, 2005, 137-142.

[8] Golombek, M. P. and Anderson, R.C. Overview of the Mars Pathfinder Mission: Launch through landing, surface operations, data sets, and science results. *Journal of Geophysical Research, 104(NO. E4)*, 1999, 8523-8553.

[9] Koratkar, A., Grosvenor, S., Jung, J., Pell, M., Matusow, D., and Bailyn, C. Science Goal Monitor - Science goal driven automation for NASA missions. Proceedings, 2004 SPIE -- Scientific Return for Astronomy through Information Technologies, Vol(5493), 2004, 33-41.

[10] Squyres, S. Roving Mars: Spirit, Opportunity, and the Exploration of the Red Planet (Hardcover). Hyperion.

[11] Steele, F. Jr., Thomas, G., and Blackmon, T. An Operator Interface for a Robot-Mounted, 3D Camera System: Project Pioneer. *IEEE Virtual Reality 1999 Conference*, 126-132.

[12] Thomas, G., J. Wagner, Z. Xiang, A. Kanduri, and J. Glasgow, Analytical Rover Operations Development, *IEEE Systems, Man and Cybernetics Conference*, The Hague, The Netherlands, 2004.

[13] Wagner, J., Thomas, G., Glasgow, J., Anderson, R.C., Cabrol, N., and Grin, E., Error-associated behaviors and error rates for robotic geology, *Human Factors and Ergonomics Society 48th Annual Meeting*, New Orleans, LA, 2004, September 20-24.

[14] Wettergreen, D., Cabrol, N., Teza, J., Tompkins, P., Urmson, C., Verma, V., Wagner, M.D., and Whittaker, W.L. First Experiments in the Robotic Investigation of Life in the Atacama Desert of Chile. *International Conference on Robotics and Automation*, IEEE, April, 2005.

[15] Wettergreen, D., Cabrol, N., Baskaran, V. , Calderon, F., Heys, S., Jonak, D., Luders, R.A., Pane, D., Smith, T., Teza, J., Tompkins, P., Villa, D., Williams, C., and Wagner, M.D. Second Experiments in the Robotic Investigation of Life in the Atacama Desert of Chile. *8th International Symposium on Artificial Intelligence*, Robotics and Automation in Space, September, 2005.

[16] Whittaker, W. and Champeny, L. Conception and development of two mobile teleoperated systems for TMI-2. In *Proceedings of the International Meeting and Topical Meeting TMI-2 Accident*, American Nuclear Society, 1988.