JOURNAL OF GEOPHYSICAL RESEARCH, VOL. 112, G04S09, doi:10.1029/2006JG000318, 2007

Comparing different methods for assessing ground truth of rover data analysis for the 2005 season of the Life in the Atacama Project

G. W. Thomas,[1] I. Ukstins Peate,[1] J. Nakamoto,[1] E. Pudenz,[1] J. Glasgow,[1] J. Bretthauer,[1] N. Cabrol,[2,3] D. Wettergreen,[4] E. Grin,[2,3] P. Coppin,[5] J. M. Dohm,[6] J. L. Piatek,[7] K. Warren-Rhodes,[2,3] A. N. Hock,[8] S. Weinstein,[9] G. Fisher,[9] G. Chong Diaz,[10] C. Cockell,[11] L. Marinangeli,[12] N. Minkley,[9] J. Moersch,[7] G. G. Ori,[12] T. Smith,[4] K. Stubb,[4] M. Wagner,[4] and A. S. Waggoner[9]

Received 28 September 2006; revised 8 March 2007; accepted 14 June 2007; published 13 September 2007.

[1] The scientific success of a remote exploration rover mission depends on the right combination of technology, teamwork and scientific insight. In order to quantitatively evaluate the success of a rover field trial, it is necessary to assess the accuracy of scientific interpretations made during the field test. This work compares three structured approaches to assessing the ground truth of scientific findings from a science team conducting a remote investigation of a locale using an autonomous rover. For the first approach, independent assessment, the daily science summaries were analyzed and reduced to a series of 1082 factual statements, which were treated as hypotheses. An independent scientist traveled to the field area to assess these hypotheses. For the second approach, guided self-study, the mission scientists themselves traveled to the field area and evaluated their own scientific interpretations. The third approach, discrepancy investigation, searched for the root causes of differences between the scientific interpretations made in the control room and those made in the field. The independent investigation provided sensitive, quantitative data, but suffered from the lack of context and continuity developed in the mission control room. The guided evaluation benefited from the context of the mission, but lacked clarity and consistency. The discrepancy investigation provided insight into the root causes behind the discrepancies, but was expensive and time consuming. The independent investigation method yielded particularly compelling results, but each method offers advantages and a comprehensive rover field trial assessment should include a combination of all three.

Citation: Thomas, G. W., et al. (2007), Comparing different methods for assessing ground truth of rover data analysis for the 2005 season of the Life in the Atacama Project, *J. Geophys. Res.*, *112*, G04S09, doi:10.1029/2006JG000318.

1. Introduction

[2] Geologic interpretations constructed from field-based studies use acquired evidence to constrain models of the possible histories and explanations related to a central hypothesis, such as the tectonic evolution of a locale. These models are ''regularly under-determined by available theory and data''; the available evidence may support multiple, equally valid, explanations. Many geologists regard the construction and selection of the preferred model as part art and part science. Nevertheless, in order to inform, guide and measure the development of technologies designed for geological investigations, such as robots for planetary exploration, it is important to evaluate the certainty, accuracy and consistency of geological interpretations.

[3] With planetary rover missions, available data is generally provided in discrete batches consisting of digital images and spectral analyses [e.g., *Arvidson et al.*, 2002; *Tunstel et al.*, 2002]. In situ field studies, in contrast, benefit from more holistic geological observations. In this respect, planetary rover missions can facilitate a critical evaluation

[1]Department of Mechanical and Industrial Engineering, University of Iowa, Iowa City, Iowa, USA.
[2]NASA Ames Research Center, Moffett Field, California, USA.
[3]SETI Institute, Mountain View, California, USA.
[4]Robotics Institute, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
[5]Eventscope, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
[6]Department of Hydrology and Water Resources, University of Arizona, Tucson, Arizona, USA.
[7]Department of Earth and Planetary Sciences, University of Tennessee, Knoxville, Tennessee, USA.
[8]Department of Earth and Space Sciences, University of California, Los Angeles, California, USA.
[9]Molecular Biosensor and Imaging Center, Carnegie Mellon University, Pittsburgh, Pennsylvania, USA.
[10]Universidad Católica del Norte, Antofagasta, Chile.
[11]British Antarctic Survey, Cambridge, UK.
[12]IRSPS, Pescara, Italy.

0148-0227/07/2006JG000318$09.00

of the quality and effectiveness of remote geological observations because of the well constrained and controlled data flow [*De Hon et al.*, 2001]. The unique interdependence of scientists and remotely operated vehicles also facilitates critical evaluation, because the scientists controlling the rover must make conscious decisions and weigh tradeoffs about where to explore next [*Backes et al.*, 2003; *Cabrol et al.*, 2007]. The critical evaluation of geological exploration with a rover facilitates study of both the process with which scientists analyze and create models of an environment and the rover instrumentation that best supports their needs [*Thomas et al.*, 2001]. The overarching aim of the research reported here is to objectively analyze the performance of a science team searching for microbial life in a remote and arid desert with an autonomous rover to better understand how to improve the process with which the science team considers the data and directs the rover, as well as to improve the design of the rover itself, including the mechanical systems, autonomous systems, instruments and data transmission strategies. This study compares three different approaches to analyzing the interpretations from the Life in the Atacama (LITA) Project (references this volume): (i) an independent evaluation of the rover-derived scientific results performed by an external scientist, (ii) a self-evaluation of the rover-derived scientific results performed by the rover science team, and (iii) a discrepancy evaluation performed by an external source focusing on three instances where alternate interpretations were observed from (i) and (ii). These different analytical methodologies allow us to objectively assess the performance of a single scientific team during a series of geologic surveys conducted with the autonomous rover Zoë [*Cabrol et al.*, 2007] in the Atacama Desert, Chile.

[4] Objectively evaluating and, ideally, quantifying the effectiveness of scientific reasoning is challenging. Geologic reasoning, which was the primary emphasis of the analysis presented in this paper, is particularly challenging both because of the complexity of geologic interpretations [*Baker*, 1999; *Frodeman*, 1995], as well as the lack of specific, overreaching constraints in interpretation scale, investigative detail and geographical limits faced by geologists operating in the field. Both scientists and philosophers have been conceptually interested in evaluating geologic reasoning [e.g., *Schumm*, 1991; *Frodeman*, 1995; *Baker*, 1999], largely because they want to understand it within the context of the broader scientific community, comparing earth or planetary scientists reasoning with the strategies employed by chemists and physicists, for example. Other scientists have compared the difference in results from satellite image mapping and field surveys, measuring accuracies of 60 to 90% [*Ferguson*, 1991; *Franklin*, 1991; *Matthews*, 1991]. Rover designers have been interested in scientific reasoning for a more pragmatic purpose – they want to know how to build better rovers. *Whittaker et al.* [1997] report that there were significant performance differences between geologists making remote data interpretation versus those using in situ studies.

[5] Field methodology of geologists has been studied in more detail, most notably by *Brodaric and Gahegan* [2001] and Brodaric et al. [2004], who evaluated geologic category development by multiple geoscientists during individual, independent field logging as part of a group project, and empirically assessed how geologists collect field data and develop geologic map units from those data, respectively. Both the 2001 and 2004 studies identified several factors that influence the process of developing categories and concepts: (i) prior knowledge (theory), (ii) data, (iii) personal bias (individuality), and (iv) experience (situation), which involves shifting physical situation and/or mental outlook. In application, this means that individuals often define the same category differently (e.g., a map unit), individuals may develop or change a category definition with time, and some categories are not fully defined by only observable data. The representational tools used by researchers are also very important [e.g., *Latour*, 1995; *Lynch and Woolgar*, 1994]. Such fluidity in geologic interpretation poses a fundamental challenge to the evaluation of geologic knowledge acquisition.

2. Background

[6] NASA has been developing autonomous rovers for planetary exploration for over a decade. A standard part of rover development is to conduct simulated missions in Mars-analogue locations. These rover field trials fall into two general categories: concept tests, which explore new equipment design, mission approaches, and exploration strategies; and field readiness tests, which ensure that a rover's equipment package is functionally prepared for a mission it is scheduled to conduct.

[7] Concept tests are common, and have been conducted with numerous rovers in a wide range of environments, including under the arctic ice [*Hine et al.*, 1994], Kilauea Volcano, Hawaii, USA [*Stoker and Hine*, 1996], Mt. Spur, Alaska, USA [*Fong et al.*, 1995; *Bares and Wettergreen*, 1997], and the Atacama Desert, Chile [*Cabrol et al.*, 2001a, 2001b]. Most tests have been in desert environments in the American Southwest, including the Painted Desert, Arizona [*Christian et al.*, 1997], the Mohave Desert, California, and Silver Lake, Arizona.

[8] Field readiness tests focused on the capabilities of the FIDO rover in preparation for the Mars Exploration Rover mission. These included the 2001 FIDO test at Black Rock Summit, Nevada, USA [*Seelos et al.*, 2001], the 2001 Mojave, California, USA field test of science operations with various anomalies [*Tunstel et al.*, 2002], and the 2002 FIDO test of Science Operations Working Group near Flagstaff, Arizona [*Tunstel et al.*, 2004].

[9] The impetus of these studies was to test-drive the rover and evaluate scientist-rover interaction. Scientific research conducted during these simulated missions emphasized evaluating the geologic history of a region, although understanding weather and, more recently, biology also played roles in these tests. The scientists' ability to perform their work is presumably influenced by several factors, including the rover's instruments, its mechanical and electrical subsystems, the information systems used to communicate between the rover and the scientists, the software interface used by the scientists, and the strategy with which the scientists organize themselves. A field test presumably allows the team to quantitatively test the effect of a new rover configuration on science productivity.

[10] This strategy is relatively effective when the subsystem being considered can be measured by some well-

defined, external metric. Reports of particular subsystems such as the rover hardware [*Tunstel et al.*, 2002], rover software, or visualization software typically focus on a performance metric such as distance traveled, absolute accuracy of the rover's self-estimate of position, the number of images downloaded and stored by the information system, or the percentage of time that the rover was autonomously controlled. For example, *Volpe* [1999] provides a careful treatment of the sun sensor, Wettergreen et al. [1997] report the performance of a vision-based control system, and *Backes et al.* [2003] describe the functioning of software used by the scientists to control the rover. Despite the complexity of the implementation and the vagaries of fieldwork, these reports enjoy a scientific validity and build a compelling argument because performance is measured on an explicit, quantitative scale.

[11] Unfortunately, the quality of the scientists' conclusions is more difficult to evaluate than a factor such as the measure of total distance traveled by the rover. Consequently it is more difficult to use science effectiveness to build a compelling argument for the benefit or disadvantage of one rover system versus another. Various authors have addressed this challenge in different ways.

[12] An overview of the last major field operations test leading up to the MER mission illustrates the gap between the stated experimental paradigm and the difficulty in drawing specific conclusions with science effectiveness as the primary dependent variable. The objectives of this mission were dominated by the information provided to the science operations work group (SOWG), which suggested that effective scientific thinking was a critical component of the test [*Tunstel et al.*, 2002].

> . . . [P]rior knowledge of the desert test site was limited to large (tens of square kilometers) aerial images and spectral data typical of real Mars orbital observations. The SOWG initially uses the aerial data to generate geological hypotheses about the field site. The rover instruments are then used to correlate hypotheses generated from these additional data to better understand the geology of the field site. This requires use of the rover's capabilities for conducting traverse science and making *in-situ* measurements in realistic terrain subject to mission-like constraints.
>
> . . . The primary objective of the operations testing was for the SOWG to use a remote rover system and rover-mounted instruments to acquire data for formulating and testing hypotheses about the geologic evolution of the field site.

[13] The paper then describes the rover, the communication and operations strategies. Surprisingly, the ''MER-FIDO Field Operations Results'' section makes no mention of any scientific hypotheses generated during the mission or their validity. It emphasizes the number of days simulated, the demonstration of different engineering performance goals for the rover, and the fact that the science team overcame various planned and unplanned operational anomalies. The only reference to the performance of the science team and their hypotheses, the stated primary objective, was that the rover traversed to locations that exhibited distinctly different geologic characteristics. Presumably, the SOWG team made geologic interpretations based on data provided by the rover, but specific interpretations, or even a quantification of the number of interpretations (successful or otherwise), were never detailed. This challenge of detailing a procedure for assessing the science productivity is common to nearly all the reports of rover field tests.

[14] Perhaps the study that most directly addresses details of science results is of a field test in the Painted Desert field experiment, Arizona [*Christian et al.*, 1997]. This emphasized evaluating rover design, but included the science objectives of establishing the general geology of the field test site as mission objectives. Other goals included simulating Pathfinder operations and implementing rapid exploration. The rover's engineering goals are described with numeric precision, such as ''onboard odometry showed the rover traveled a total of 469 m,'' while the science results are presented qualitatively. For example, ''comparison of half and full resolution images shows that many conclusions cannot be made with the lower resolution images alone,'' ''locating the rover position and orientation from imaging in the descent photos was critical and often difficult,'' and ''in one anecdotal view, the Marsokhod 'missed 90%' of the interesting geology it encountered.'' This last quote is tantalizingly close to the numeric precision that would allow a system to be objectively evaluated; however, the anecdote seems to be a figure of speech rather than a systematic, objective, and repeatable evaluation of science effectiveness. An attempt at quantification is also illustrated in Hayati et al. [1997] when they state, ''The tests showed that remote geology is indeed possible using an *in-situ* robotic vehicle with the science team successfully identifying the geology of the site to within an 80 % accuracy.'' Unfortunately the method used to derive this metric is not provided. A percentage measure of science accuracy implies some enumeration of the amount of science gotten right divided by the total amount of science proposed. Both the enumeration and the determination of correct and incorrect pose interesting challenges.

[15] Several other issues arise from rover testing protocols. First, there is a difference between evaluating that instrumentation is functioning to a standard set of defined operating parameters, versus optimizing the instrumentation to maximize its effectiveness in providing scientists with the most useful data to make the necessary observations, as defined by the mission objectives (i.e., finding life, evaluating the presence or past presence of water). Also, field tests on Earth provide the only opportunity to verify scientific observations, to quantitatively compare interpretation of restricted rover-based remotely derived data with in situ interpretation in which visual data is unrestricted and unlimited, in order to identify and evaluate the nature of and reason for any differences between the two data sets on the same locale. Any differences may reflect subtle issues such as the restricted type of available data, generalized versus specific instrumentation types, lack of tactile input, image resolution, or even geoscientist interface with the data, which may not become apparent without such a quantitative comparison. Given studies that clearly demonstrate interpersonal differences in methods of data analysis and interpretation, it is likely that rover-based geology requires a configuration of skills and techniques that make it a unique specialization. This is an opportunity to evaluate the influence of background knowledge on remote data interpreta-

tion, and also to train and enhance performance in what is no doubt a highly specialized and challenging field.

[16] Recent studies that directly addressed these concerns evaluated the performance of a geoscientific team analyzing data from a simulated rover in the Mojave Desert, Grey Mountain, Arizona, USA [*Thomas et al.*, 2004; *Wagner et al.*, 2004], in order to directly target and identify the nature of, and reason for, any discrepancies between the geoscientists' understanding of rover-derived data compared with field-derived observations from the same scientists at the same locale. Three geologists participated in a simulated rover mission, making observations and analyzing data similar to that which would be collected by a rover, but which was systematically gathered by hand at a remote location. After the analysis was complete and the geologists formulated a conceptual understanding of the geologic history of the area from the rover data, they were led, blindfolded, to the location of the test site. They were then encouraged to describe the differences between what they had understood about the environment based on the rover data and what they observed in situ, which resulted in identification of twenty differences in interpretation between the two. Discrepancies included underestimating the prevalence of a lithology (basalt) and failing to find the clear evidence for a streambed in exposed rock.

[17] The transcripts of the geologists' conversations in the control room and the field were analyzed. Each specific scientific utterance representing an observation, interpretation or hypothesis was individually coded, and the observations and interpretations supporting each hypothesis were enumerated [*Wagner et al.*, 2004]. Based on a model of human error proposed by *Rasmussen* [1983], the scientific utterances were categorized as skill, rule and knowledge based decisions, and the knowledge-based decisions were further analyzed for particular patterns of error. While the specific findings of that work are beyond the scope of our paper, this novel method is promising as a technique for enumerating the amount of science produced (individual scientific utterances) and for determining the accuracy of the utterances by having the scientists themselves determine this in situ. Quantifying the amount of knowledge is clearly problematic because knowledge is not a substance. However, it is quite common for scientists to evaluate the success of a robot mission with phrases like, "we did a lot of good science today." If this is the manner in which a rover field test will be evaluated, it is necessary to understand the metric to which the scientists are referring.

[18] The study also highlighted difficulties with this approach, in that verbal utterances provide a rather informal mechanism to represent scientific thought [*Wagner et al.*, 2004]. There is a rich literature on human problem solving based on verbal protocol, but the success of the analysis often lies in the clarity and structure with which the experimenter can model the logical paths that the test participants might follow [e.g., *Ericcson and Simon*, 1984]. Unfortunately, in the experimental setting considered here, we do not yet understand the structure of the scientists' reasoning process in enough detail to recognize which logical path they are following and what other paths they are not following. Sometimes important evidence is not mentioned because it seems apparent, while at other times scientists make unsubstantiated speculations that they would hesitate to put in print. Counting written statements in a formalized report prepared by the scientists would overcome these challenges. Another problem encountered is the lack of an absolute scale of truth, or "correctness". If a scientist is unaware of a phenomenon or pattern he or she may miss this in the rover data and in the field. However, the experiment simply neglects this potentially valuable information rather than measuring it as missed information. One way to avoid this source of error is to have many scientists participate, in order to reduce the chance of a smaller, less diverse group overlooking an important fact or observation. Finally, while the counting of utterances provides a reasonably repeatable, relative scale, it does not provide for a means to rank the importance of observations.

3. Three Assessment Methods

[19] This study analyzed the performance of 5 scientists participating in the 2005 season of the LITA Mission project. Details of the scientific mission, instrumentation used on the rover, and scientific interpretations based on the rover-derived data can be found in other works in this special issue [*Cabrol et al.*, 2007; *Warren-Rhodes et al.*, 2007] (J. L. Piatek et al., Surface and subsurface composition of the LITA 2004 field sites from rover data and orbital image analysis, submitted to *Journal of Geophysical Research,* 2007, hereinafter referred to as Piatek et al., submitted manuscript, 2007; P. E. Coppin et al., Life in the Atacama: Remote science investigation tools and human performance, submitted to *Journal of Geophysical Research,* 2007; S. Weinstein et al., Application of pulsed-excitation fluorescence imager for daylight detection of sparse life in tests in the Atacama Desert, submitted to *Journal of Geophysical Research*, 2007, hereinafter referred to as Weinstein et al., submitted manuscript, 2007). Over the course of the mission, scientists' activities were monitored in the control room (Carnegie Mellon University, Pittsburg, PA, USA) with both audio and visual recordings. Cameras and microphones were placed in the control room and common meeting areas, and each scientist wore a lapel microphone. In addition, the scientists were also monitored by at least two observers throughout the experiment. These observers collected a variety of notes, including a narrative description of the activities of each scientist every five minutes, and classified the observed activities into one of 19 categories (for further details, see *Pudenz* [2006] and *Pudenz et al.* [2006]).

[20] The science team varied slightly throughout the period studied, but generally included at least three geologists and two biologists. The team followed a daily schedule that began with the download of a new data set from Zoë at around 7 pm each day of the mission. They would analyze this data and refine their plan for Zoë's activities the next day, and by the end of each working day, typically between 1 and 3 am, the science team constructed and uploaded the next day's command sequence back to Zoë. This command sequence included a list of locales to be investigated along a traverse, desired data collection activities, and a priority score for each activity. Later that morning, typically around 11 am, the team would reassemble to further analyze the previous day's data and prepare their daily science summary. In the afternoon, the scientists met as a

group to collaborate on interpretations and construct a preliminary plan for the following day's activities. Around 7 pm the next download arrived from the rover and the cycle repeated for each day of the mission.

[21] In addition to monitoring and observing the scientists during the mission, all archived mission data, including images, spectra, and other data collected by the rover, the command sequences sent to the rover, reports on the success of each day's command sequence created by the field team running the robot, and the daily reports, annotated images and maps created by the scientists were compiled for this study.

[22] The following sections describe three different approaches to evaluating the accuracy of science generated during the LITA mission. The first section describes an independent assessment of the rover-derived data. This is followed by a description of the guided self-assessment completed by the rover science team-members themselves. Finally, a discrepancy investigation is presented on three incidents of importance to the mission.

3.1. Independent Assessment

[23] The LITA mission scientists conducted a detailed study of the geology and the existence of microbial life, and life-sustaining habitats, in the Atacama Desert, utilizing the rover science instruments payload (detailed in *Cabrol et al.* [2007] and Weinstein et al. (submitted manuscript, 2007)). Because of the specialized nature of many of the biology-based analyses, and the difficulty of verifying the presence of chemical signs of life such as peptides, sugars and DNA that played a key role in many biological experiments, this study chose to focus more closely on the geological and broad-scale biological scientific observations and interpretations, because it was thought that these would be more straightforward to identify and evaluate. For the independent assessment, specific scientific statements were garnered from the science activities in the control room, and the accuracy of those statements were tested by an independent geologist in the field. Because of the informal nature of verbal utterances, we chose to focus on the scientists' daily summaries in which they presented a more formal version of their hypotheses based on the most current rover data to which they had access.

3.1.1. Method

[24] Thomas and Ukstins Peate parsed the science daily summaries completed over 21 exploration days into individual statements representing observations, interpretations and hypothesis proposed by the scientists, making note of the context and origin of each statement (scientist proposing, data upon which it was based, locale or site where data was collected). This parsing resulted in 1082 specific statements that could be independently evaluated as being either true or false (see auxiliary material[1] for a complete list of hypotheses). One significant reason for this parsing was to separate out discrete information packets found within complex and multi-component statements. For example, the following sentence, taken directly from the science notes, was parsed into the statements (hereafter called hypotheses) presented below:

[25] Statement: The regional slope is interrupted by clusters of promontories, local topographic basins, and drainages, some of which appear to be controlled by major northeast- and northwest-trending tectonic structures, and materials of varying morphologic characteristics and albedo.

[26] Parsed hypotheses: Regional slope is interrupted by clusters of promontories. Regional slope is interrupted by local topographic basins and drainages. Some of the basins and drainages appear to be controlled by major northeast- and northwest-trending tectonic structures. Region has materials of varying morphologic characteristics and albedo.

[27] The parsing of scientists' reports into individual, true or false hypotheses allows for a quantitative evaluation of the effectiveness of the science completed during the mission. One step in evaluating this was to re-visit 59 locales visited by the rover, with the parsed statements relevant to each locale, to independently evaluate whether each hypothesis could be confirmed, refuted or was unverifiable at that time.

[28] Thomas and Ukstins Peate conducted this verification in October 2005, at the close of the LITA field campaign. Thomas, who has a background in human factors research, observed, but did not participate, in the scientific analysis during the mission. Ukstins Peate, a professional geologist, did not participate in the rover mission and provided an independent field judgment of the statements and observations made by the science team during the mission. In many cases, some judgment was required to interpret the meaning or intent of the original statement, such as ''rocks are dark colored'', because the nature of the hypothesis was not quantitative. A total of 969 hypotheses were confirmed or refuted in the field, with an additional 113 classified as unverifiable because they contained interpretations that were difficult or impossible to refute in the field or with the available resources. Unverifiable hypotheses were either related to time-dependent phenomena, such as ''no visible moisture,'' or ''no cloud cover'', or were unilaterally true and therefore not disprovable, such as ''The history of fan emplacement includes environmental change'', or ''Unit F has various sequences of emplacement''.

[29] After field assessment was complete, the statements were divided into eleven categories based on the type of statement (material identification, position estimation, material diversity, material color, geomorphology, moisture presence, presence of life, material size, material pattern, material shape and material source) and further subdivided into either observation or interpretation. Some categories, like material size, consisted of only observation-level statements. Other categories, such as geomorphology, consisted of both observations and interpretations. The statements were further analyzed to evaluate any statistical trends in data source, such as which instrument provided the data upon which the statement was based.

3.1.2. Results

[30] Of the 969 hypotheses that were tested, the opinion of the independent geologists differed from the science hypotheses 202 times (~20%). Table 1 presents the distribution of the hypotheses into categories and statistical information on evaluation of those hypotheses.

[1] Auxiliary materials are available at ftp://ftp.agu.org/apend/jg/2006jg000318.

Table 1. Geological Hypothesis Accuracy by Hypothesis Category

		Geomorphology[a]	Material Size	Material Color	Material Source	Material Diversity	Moisture Presence[a]	Material Identification	Position Estimation	Material Pattern	Presence of Life	Material Shape
Interpretation	Total hypoth.	52	NA	NA	41	84	63	128	9	NA	NA	NA
	Differing hypoth.	14	NA	NA	1	23	9	57	5	NA	NA	NA
	% Differing	27	NA	NA	2	27	14	45	56	NA	NA	NA
Observation	Total hypoth.	202	136	120	NA	NA	20	NA	16	16	16	67
	Differing hypoth.	36	16	26	NA	NA	4	NA	6	1	1	4
	% Differing	18	12	22	NA	NA	20	NA	38	6	6	6
Interp. and obs. combined	Total hypoth.	254	136	120	41	84	83	128	25	16	16	67
	Differing hypoth.	50	16	26	1	23	13	57	11	1	1	4
	% Differing	20	12	22	2	27	16	45	44	6	6	6

[a]The geomorphology and moisture presence categories separate hypotheses related to interpretation from those related to observation.

[31] Of the 392 hypotheses categorized as interpretations, the independent scientist differed on 111, or 28 %. Of the 577 hypotheses classified as observations, the independent scientist differed on 93, or 16 %. The independent scientist differed with 51 of the 209 hypotheses traced to the biologists (e.g., those observations made by the biologist that the independent scientist felt qualified to assess), or 24 %, and differed with 148 of the 737 hypotheses traced to the geologists, or 20 %. Images from the workspace camera, the fluorescence imager (FI), and the stereo panoramic imager (SPI) generated 26, 440, and 325 hypotheses each, spectral data from the TIR and VIS/NIR generated 14 and 15 hypotheses, respectively, and satellite images produced 155 hypotheses. The satellite, FI and SPI images were the only data types with enough samples to produce statistically reliable conclusions and the independent scientist disagreed with 26 %, 19 % and 18 %, respectively, of hypotheses based on these data.

3.1.3. Discussion

[32] The results suggest that hypotheses that were categorized as observations were more frequently correct than those classified as interpretations (16 % versus 28 %). It is interesting that the traditional formula for the reliability of a process (referred to as R1) that depends on two steps (referred to as R2 and R3), each with a reliability of from 1 to 16 % (based on the overall reliability of observations, Table 1) yields an expected failure rate for R1 of 1 – R2*R3, or 29.5 %, tantalizing close to the value of 28 % observed in this experiment [*O'Connor*, 2002]. In this case, the reliability values depend on the subjective process of identifying the statements in the science reports, classifying the statements into predefined categories, and independently evaluating their accuracy. The mathematical computation also depends on the assumption that two observations are required for a single interpretation, which may not necessarily be the case. If a correct interpretation needs to be based on more than one correct observation, then interpretations would be less reliable than observations.

[33] The results also suggest that the discrepancy rate between the independent scientist and the rover mission scientists is reasonably consistent at 20 to 24 %, irrespective of rover mission scientist focus field. This value should not be interpreted as an accuracy rate for the professionals, since in some cases there were legitimate differences of opinion for items marked as a discrepancy. Often these differences of opinion were the result of differences in the interpretation of and meaning of words. For example, the biologists' and the independent geologist's definition of desert pavement differed because the biologists used the term to describe a habitat consisting of loosely scattered rocks with volumetrically small amounts of visible fine-grained sediment, whereas the geologist interpreted the word to mean that the ground surface had been deflated and the rocks were closely packed with the appearance of a jigsaw puzzle.

[34] The analysis of the data sources for the hypotheses revealed that the satellite images were a somewhat less reliable source of hypotheses (26 %) than the FI images (19 %) and SPI images (18 %), when compared with the independent geologist findings in the field. This result supports the importance of ground reconnaissance for planetary missions as a means to confirm findings from satellite images. It is also clear, however, that neither instrument provides a source for completely reliable scientific statements. This may be a function of the type of statements that scientists were interested in making rather than a function of the instrument itself. For example, the scientists may avoid making remarks that are obvious to them based on the data, and instead focus their hypotheses on information that required more complex, and consequently more risky, interpretations.

[35] An analysis of the rate of discrepancy in the hypotheses from different categories revealed that the categories could be divided into three natural groupings based on discrepancy rates. The least risky category contains hypotheses about material shape, size and source. The scientists in the control room and the independent geologist agreed on these hypotheses between 92 and 98 % of the time. The scientists in the control room and the independent scientist in the field agreed on hypotheses about material diversity, pattern, color, geomorphology, and the presence of moisture or life between 70 and 90 % of the time. They only agreed on hypotheses that identified the material composition or the location of the rover between 50 and 60 % of the time.

[36] This method of evaluation provides a second, independent data set for comparing geological observation and interpretation of a field site and specifically addresses the exact observations and interpretations made by mission scientists and their context within written reports. Although this provides a high level of detail and addresses individualized questions that can be tailored to each mission, it suffers from a lack of context as developed by the mission science team. These groups often collaborate over multi-year periods and develop a group working methodology and shared understanding which may influence or provide unspoken nuances to their observations and interpretations,

but is generally not documented and therefore is inaccessible to the independent geologist.

3.2. Guided Self-Evaluation

[37] The second assessment technique relied on the scientists' evaluation of their own conclusions. This method has potential for deepening the science team's understanding of their working methodology, as well as the science results themselves, though direct evaluation of their own observations and interpretations carried out *post-hoc* at the remote locale. For some science studies conducted as a part of a rover field test, additional field study can play a key component in enhancing discovery [e.g., *Warren-Rhodes et al.*, 2007]. This evaluation methodology may be a way to reflect on the analysis methods and results in order to develop recommendations that would improve future rover-based scientific studies.

3.2.1. Method

[38] Because of time constraints and logistical concerns, it was not possible to have the whole team go to each site visited by the rover and re-evaluate every point made in the science reports. Thus, a subset of sample locales were selected because they held particular scientific interest, either because the locale was central to the scientists' interpretation or because it contained important features characteristic of the region. For each of the seven target locales, Thomas and Ukstins Peate developed a list of 3 to 9 questions reflecting the key scientific hypotheses found at that locale (see auxiliary material for complete list of questions). These questions were posed in the style of the parsed hypotheses, as one true or false statement containing a single idea or concept. In many cases the question was a simple rephrasing of the original hypothesis put forth by the scientists, sometimes rewritten to illicit the opposite 'answer' (i.e., false rather than true), while other questions were constructed to draw the scientists' attention to an important feature of the environment that had been missed in the original analysis. In this way we hoped to evaluate not only the effectiveness of the documented science, but also to begin to address the significance of potentially valuable information that has been missed. For example, the following statements were included in the survey: ''Rocks [at this site] are mostly white,'' ''Landing site characterized by small basins/playa,'' ''White rocks are ash.''

3.2.2. Results

[39] In January 2006 four of the mission scientists and investigative team, consisting of Ukstins Peate, Thomas, Pudenz and Glasgow, traveled as part of a larger group including the mission managers, rover engineers, and other observers to each of the seven locales. At each locale the scientists were given a copy of their original reports and the prepared question list, which contained a total of 49 questions. The scientists were then asked to individually indicate whether the statement was true or false. Three geologists (including Ukstins Peate) and two biologists participated in the experiment.

[40] Not all of the scientists chose to respond to every question. All five scientists answered the same ten questions. Twenty additional questions were answered by four of the scientists. Eleven more questions were answered by three scientists, seven were answered by two scientists and just one was answered by one scientist, totaling 49 questions. In many cases the scientists chose to elaborate on their answers by adding comments in the space provided. Often these comments addressed their lack of clarity in understanding the question and noted the need for interpretation on their part. The mission geologists responded to between 34 and 46 questions each, while the mission biologists responded to between 29 to 32 questions each.

[41] For questions that more than one scientist responded to, there was 100 % agreement for 24 out of 48 questions, or 50 % of the time. The most common controversy (occurring in 10 of the 24 questions with disagreement among answering scientists) involved the independent evaluator disagreeing with the other scientists. Including both controversial and uncontroversial statements, this pattern characterizes 20 % of the total statement responses. Five other controversies involved one respondent disagreeing with two others; these dissenting voices were evenly split among the rover mission scientists.

[42] Observations as well as interpretations were sometimes controversial. Some of the controversial statements were direct observations, such as: ''fractures are visible in the white, crusty material''; ''clasts are light and dark colored''; and ''wind-blown materials partly mantle surface''. Some of the controversial statements required interpretations: ''conglomerates are present'' and ''clasts display desert varnish''. Those comments for which there were agreement were also balanced between direct observation, such as ''rocks are mostly white,'' ''plants are present,'' and ''white rocks display black spots,'' and those statements that require interpretations, such as ''black rocks are volcanic'' and ''the landing site has a pervasive anhydrite subsurface layer.''

3.2.3. Discussion

[43] The frequent lack of agreement (50% of 48 responses) among the scientists suggests that a simple statement may not contain enough information to form an unambiguous conclusion. In many cases the comments indicated that there were reasonable differences of opinion about what the statement meant. To some extent, however, this ambiguity may be traced back to ambiguity of language in the original hypothesis derived from the daily science reports. Consequently, it seems likely that the context in which the statement was originally made should be carried through to the evaluation of the information.

[44] The disagreement between the independent geologist and the mission scientists may be a consequence of this lack of context. The mission scientists, working over a period of three years, naturally developed an understanding of what they really meant when they used particular phrases, such as desert varnish or desert pavement, as documented by *Brodaric and Gahegan* [2001]. This understanding of language that evolved over the course of the mission was not shared by the independent geologist, who had to rely on other sources of context for the meaning of these terms.

[45] Deciding on the best locales to visit was at first a source of anxiety. The science team was concerned that the needs of this assessment would divert them from the areas in which they were most interested. The concern vanished when both the science team and the investigators compared their list of desired target sites and saw that there was excellent agreement in which would be the most interesting

places to visit. Such concurrence was not evident with the selection of questions, however. Although there was little direct evidence for this point, it appeared that the scientists found the need to reflect on the list of hypothesis statements to be a burden, and they preferred to pursue their own ideas in the field. This may have been a consequence of misestimating which would be the best hypotheses to consider, or it may be a consequence of the scientists' enthusiasm to explore the larger test area.

[46] Although this assessment technique provided interesting insight into the divergence of viewpoints within the science team and the contextual meaning of scientific terms, it was less effective in providing compelling insights into the design of future rovers or adjustments to mission procedures.

3.3. Discrepancy Investigation

[47] The third assessment approach involved conducting a thorough, *post-hoc* investigation of three critical discrepancies observed between the work in the control room and the observations in the field. The three discrepancies were selected because they appeared to directly address problems with the instruments, analysis techniques, procedures or rover design. The purpose of the analysis was to reveal the root causes of each critical discrepancy [*Nakamoto*, 2006]. The structure of the analysis was based on the methodology used to investigate industrial accidents [*Vincoli*, 1994], traffic accidents [*Wheat*, 2005], and airline incidents [*Walters and Sumwalt*, 2000]. The first discrepancy centered on an annotation on a geological map created in the control room that a particular unit observed on the satellite image was volcanic, although in the field the region was characterized by a light layer of dust on a salt flat, or salar. The second discrepancy centered on a mention in the science report that the rocks at a particular site "are mostly white," although, in the field, white rocks appeared to be rare. The third discrepancy was a surprisingly large position estimation error that occurred over a period of several days during the investigation of a site.

3.3.1. Method

[48] The Discrepancy Analysis Protocol described by *Nakamoto* [2006] was applied by an engineering graduate student who had not participated in the control room observations or fieldwork. The principal steps in the protocol are: (1) prepare for an investigation, (2) identify discrepancies, (3) gather evidence, (4) run analyses, (5) determine root cause(s), (6) develop solutions and recommendations, and (7) implement and monitor changes to the system. Steps 3 and 4 may be repeated as often as necessary. The investigator was instructed to use the video footage, audio recordings, science reports, samples collected from the field, and the information on the science website as the source material for the investigation. All findings and conclusions needed to be supported by these materials or eyewitness testimony.

[49] In the interest of space, the report of all three discrepancies are not reproduced here. Instead the main points observed for the first discrepancy are reported in the results ection. The two remaining discrepancy reports can be found in *Nakamoto* [2006].

3.3.2. Results

[50] This discrepancy investigation focused on a specific annotation in the science summary, which stated: "Unit Pmb is volcanic based on spectral analysis." Unit Pmb was identified during ground truth evaluation as a salar, and when this statement was presented to the science team during guided self-evaluation, three scientists marked it as false, The others noted that the statement was only partially correct and that other issues should be considered. This discrepancy was selected for investigation because it was one of the rare instances in which a control room hypothesis was overturned in the field.

[51] The origin of this comment appears to come from a conversation about Unit Pvb between the spectroscopist and two of the biologists. The conversation was triggered by analysis of the satellite data covering the potential landing ellipse area, which included the orbital image, the Advanced Spaceborne Thermal Emissions and Reflection Radiometer (ASTER) data, and the IKONOS data. Using the satellite data, the spectroscopist ran preliminary analyses of the landing ellipse prior to the start of the rover mission to determine areas of interest based on morphology and spectral composition. The reflectivity analysis suggested that Unit Pvb was a salar.

[52] Next, the spectroscopist conducted several analyses of the visible infrared and thermal infrared spectra of the region that suggested that Unit Pvb contained volcanic-derived material (Piatek et al., submitted manuscript, 2007). A conversation between two biologists and the spectroscopist occurred at 13:45 on Sol 8, and revealed that salt had no spectrum; however, according to the TIR, the salar areas showed a non-flat spectrum.

[53] At 15:55 the spectroscopist presented the spectral analyses results to the entire science team. During this meeting, the spectroscopist expressed interest in the areas labeled salar because of the discrepancy between the TIR and reflectivity findings. The next day, a summary of the previous day's findings was included in a presentation. This presentation used similar interpretations for both unit Pmb and Pvb. It suggested that the science team hesitated to conclude that Unit Pmb was volcanic-based.

[54] Ironically, as Piatek et al. (submitted manuscript, 2007) explains, Pvb turned out to be a mudflat rather than a salar, but Pmb was a salar covered with a thin layer of dust. A major-element X-Ray Fluorescence analysis of both the salt (semiquantitative) and dust coating from Pmb is presented in Table 2. Due to the high silica concentration, the dust is a mixture of quartz and aluminosilicate-rich clays that may have been derived from surrounding volcanic and sedimentary sources.

[55] The science team had formed an hypothesis encompassing both units Pmb and Pvb, which was correct for Pmb but incorrect for Pvb. However, the written record of this incident was annotated "Unit Pmb – Plains-forming materials of medium albedo (tan). *Interpretation*: Volcanics based from spectral analysis using ASTER, though little confidence in this interpretation." This statement was included in the list of testable statements as "Unit Pmb has volcanics (based from spectral analysis, low confidence)." Then, the statement "Unit Pmb is volcanic based on spectral analysis" was included in the questionnaire which was given to the scientists in the field. Ultimately,

Table 2. Chemical Composition of Samples Collected in the Field

Sample	SiO_2	TiO_2	Al_2O_3	Fe_2O_3	MnO	MgO	CaO	Na_2O	K_2O	P_2O_5	Total	LOI
Pmb salt[a]	41.8	bdl	7.0	0.7	bdl	0.7	341.0	291.0	0.1	bdl	72.6	27.4
Pmb dust	67.7	0.3	10.4	2.3	0.1	1.3	3.4	2.9	2.6	0.1	91.0	8.8

[a]Semiquantitative analysis; bdl, below detection limit.

the scientists' potential interpretation of Pmb as a salar was lost in their daily summary, upon which the ground truthing was organized.

[56] In order to organize the evidence in a systematic and chronological order, a causal flow chart (Figure 1) was created as the evidence was gathered. This chart illustrated all relevant events that contributed to the discrepancy. The major contributors, or causal factors, were identified in these charts and investigated further. Each causal factor was further developed in a root cause map (Figure 2). Root cause maps help investigators identify the underlying causes for a specific causal factor. The root causes, which are identified as factors that might be controlled or changed in the future are then attributed to the discrepancy as a whole.

[57] From the completed hypothesis causal factor chart (Figure 1), two causal factors were identified: thermal infrared (TIR) indicates that Pmb is volcanic, and the reflectivity analysis indicates that Pmb is a salar. Using the causal factor chart, a single root cause map (Figure 2) was created to determine the root causes of the discrepancy.

[58] The investigation of the hypothesis ultimately found two root causes for the discrepancy: (i) Zoë did not visit unit Pmb, and (ii) the documentation of Pmb did not adequately describe the scientists' understanding of what that unit might contain. The only documentation was the annotation of the orbital image, which failed to mention that the scientists were considering a salar as a viable alternative hypothesis.

3.3.3. Discussion

[59] The absence of any documentation on the science team's low confidence in their hypothesis is problematic because it becomes more difficult, for all parties, to accurately determine the reasoning behind such a rating. During a conversation with two biologists the spectroscopist hypothesized that Unit Pmv (and by extension, unit Pmb) could be a salar covered in volcanic dust from the mountains. With the original hypothesis from Table 1, there is still the possibility that the mantle covering the salar has volcanic components. Thus, Unit Pmb could be a salar and still return a volcanic signature in the TIR analyses. The spectroscopist's undocumented hypothesis suggests that Unit Pmb is a salar covered with volcanic dust, which was a more accurate interpretation. Other than the two biologists in the conversation, there is nothing to suggest that the rest of the science team heard the spectroscopist's original hypothesis. Since it was not documented, this hypothesis would have been lost if this investigation had not been conducted.

[60] One of the root causes of the discrepancy was that Zöe did not visit Unit Pmb during the field test. In this case, a rover investigation of the site would almost certainly have revealed the satellite misinterpretation, as the investigation of Pvb corrected the error for that locale. Unfortunately, the two interesting regions were in opposite directions. The fact that the scientists recorded their low confidence, coupled with their success with Zöe's investigation of Pvb, suggests that the analysis of the region from the different scales afforded by combining satellite and rover data helped the

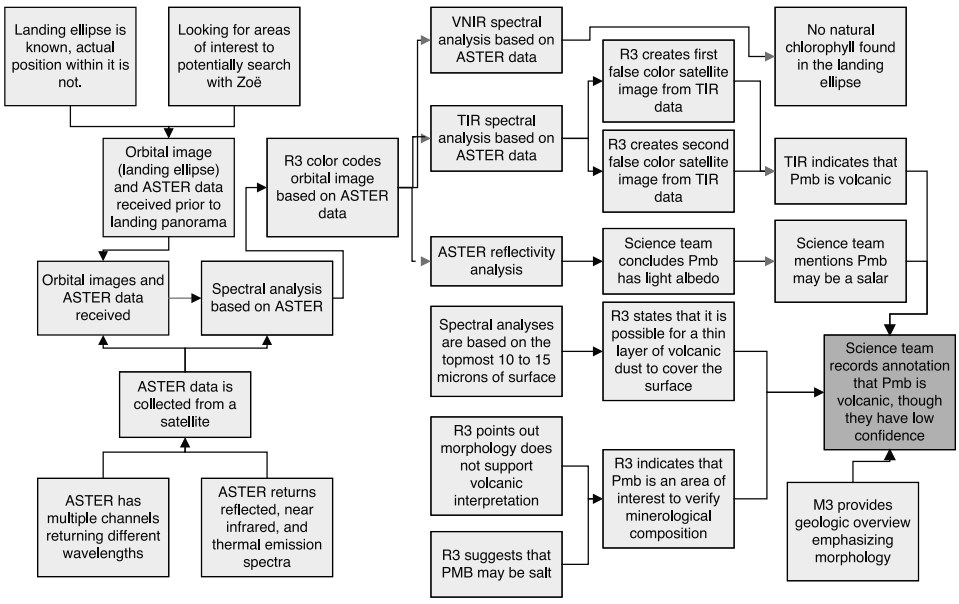

Figure 1. Causal factor chart. M3 refers to a geologist. R3 refers to a spectrologist. The primary causal factor is indicated by the dark box.

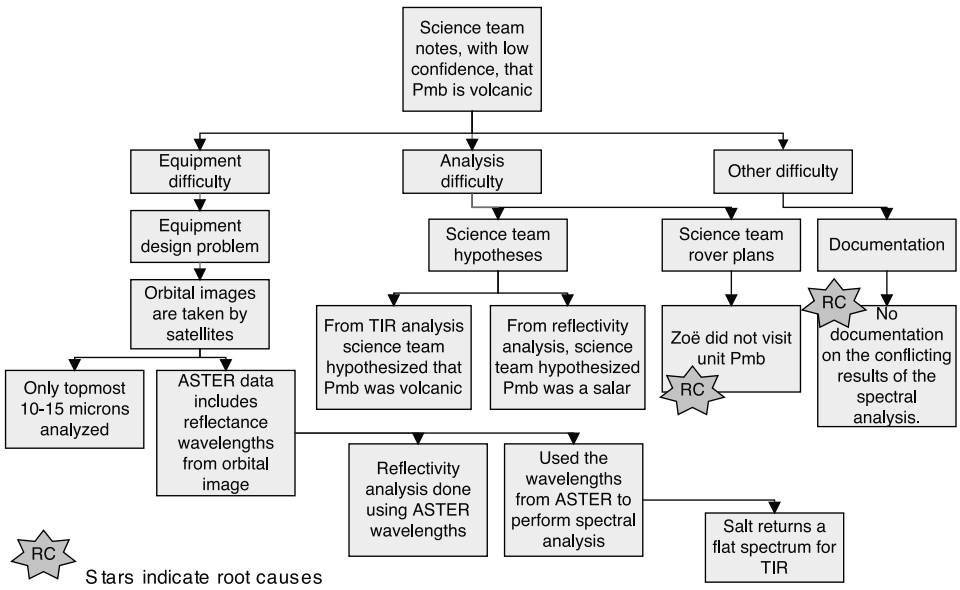

Figure 2. Root cause map.

scientists to fully understand the strengths and limitations of their study [*Fink et al.*, 2005]. Thus, the Pmb discrepancy may be attributed to resource limitations that forced the team to choose between the two potentially interesting targets; their process in using the rover seemed to be capable of resolving this discrepancy had they chosen to pursue it. Indeed, Piatek et al. (submitted manuscript, 2007) also conclude that spectral diversity observed from orbit does not necessarily accurately reflect the diversity of compositions at the surface, and therefore rover-based studies are necessary in order to augment the limitations of orbital data sets.

4. Conclusions

[61] The application of three different assessment techniques to evaluate the same rover field test provides a unique opportunity to understand and explore the strengths and weaknesses of each technique in quantifying the success of remote exploration. The independent assessment provides clear quantitative performance metrics that are helpful for tracking changes from one test or field site to another, and between field tests. The metrics may allow the scientists, engineers and mission managers to clearly distinguish sources of errors and trace these back to people, events and instruments. The clarity of the numbers, however, hides the subjective, interpretive process with which the values are derived. Scientists with different working styles could provide very different values for the same experimental conditions. A scientist who records obvious observations of each image would tend to accumulate fewer discrepancies than a scientist who only states 'interesting' or risky interpretations. Also, multiple independent assessors would likely come to different conclusions about various testable field hypotheses, particularly given the independent assessor's lack of context shared by the original science team.

[62] If statement-by-statement assessment of daily science reports became the standard technique for evaluating rover field trials, science teams might begin to alter their behavior to make the process of assessment more consistent and reliable, at least for observations and interpretations that are easily testable. This could allow low-level perceptual problems to be identified and addressed by future missions. To avoid the risk of being misinterpreted by the field assessor, science teams might try to standardize their language. If, over time, an interdisciplinary group develops new words or uses existing technical words but shifts their meaning, it would be necessary for them to clarify these meanings with independent scientists performing ground truth tests.

[63] The clarification process might reveal and provide the opportunity to resolve differences in understanding within the science team. This process ensures that the meaning of the science documentation is not so dependent on context that a competent outsider would not be able to accurately interpret its meaning. In order to control the types of judgments that would be used to assess performance, the scientists might evolve a standardized interpretation reporting style. To ensure that interpretations were fairly confirmed or refuted, the scientists and assessors would be motivated to develop a common understanding of how the interpretations should be tested in the field. In this sense, external assessment of performance would help to develop a level of rigor that would ultimately support science that is more clear, precise and repeatable. The risk of such standardization is that it would tend to deemphasize creativity in interpretation. Strict protocols are more effective at predictable, standard work practices than they are for creative practices. Although some aspects of geology, such as the accepted cues to defining rock type, are relatively consistent, others, such as developing a comprehensive geologic history, are not formally structured by a single set path. Thus, strict set protocols would be more applicable to some aspects of geologic interpretation than others, and a multicomponent non-linear protocol, capable of being modified by the user, would be more applicable to other interpretative aspects.

[64] The guided self-assessment approach allows for greater flexibility in assessment, but produces less compel-

ling results for outside observers. This approach has the advantage that the scientists can focus on the items of immediate interest to their investigation without worrying about the need to standardize their findings to outsiders until after they have finished forming their opinions. The self-evaluation process allows the science team to emphasize developing a shared understanding of the remote environment and decreases the controversy associated with poor contextual understanding. Deferring any needed clarification of context-dependent information until after the final evaluation provides the opportunity for the meaning of words and ideas to subtly shift so that differences between what was believed in the control room before the field visit and what is believed after the field visit are lost in the shifting fog of incompletely defined language. However, this context-dependent understanding may not translate well to the outside community and may limit how the experience of one mission may be communicated beyond the immediate group. Consequently, the designers and mission managers may become dependent on a group of highly specialized mission scientists to provide a heuristic assessment of the benefit of new technology.

[65] The discrepancy investigation assessment can provide excellent detail and reveal the history of observation and interpretation leading up to the final product, which is missed by the above two methods, particularly for contextual items that are easily overlooked with the independent evaluation. Unfortunately, the discrepancy evaluation is the most time consuming and expensive technique, because many hours of audiotape and video footage must be reviewed to resolve each issue. The discrepancy investigation approach is perhaps the most powerful approach for forcing an investigator to reconsider initial biases, because it emphasizes the entire context in which an idea was developed. Investigating these after the fact can provide insight that was not initially available. Because of its expense, however, this technique might be best reserved for those events that the science team and mission managers deem to be both of the greatest interest to the team and that hold the greatest potential for future insight into design requirements for future mission protocols, rovers and instruments.

[66] Each of the three approaches plays an important role in assessing a rover field test and is applicable to future remote exploration missions. If only one method could be selected, the independent observer approach offers the greatest promise because of its broad evaluation, clarity and the positive dynamics it could create for the research field. The other two approaches, however, have important insights to offer and can be applied more selectively to assess well-defined discrepancies deemed to be of great significance to the outcome of the mission. The self-evaluation protocol provides great opportunities to make unexpected discoveries about the impressions formed by the science team. The discrepancy investigation approach may have the greatest potential to overcome subtle biases and shared interpretations built up by a science team or the observers who watch them. Also, self-evaluation and discrepancy investigation processes are not limited to future missions, but can be applied to current Mars MER missions if a situation arises that necessitates such critical analysis. Much could be learned about the current workings of remote exploration science methodology that would ultimately enhance and inform future exploration. Ultimately, the most reliable assessment would involve a combination of all three methods conducted with the knowledge and encouragement of the science team from the outset of the mission.

[67] **Acknowledgments.** We are grateful to the science team for their patience and assistance with this effort throughout the LITA project. We would like to thank the Obermann Center, University of Iowa, for providing G. Thomas and Ukstins Peate with Interdisciplinary Scholar awards in June 2006, where the philosophical underpinnings of this manuscript were developed. This work is partially supported by NASA's Applied Information Systems Research Program under NASA grants NAGW5-11981 and NNG05GA51G and the Astrobiology Science and Technology for Exploring Planets under grant NAG5-12890.

References

Arvidson, R. E., S. W. Squyres, E. T. Baumgartner, P. S. Schenker, C. S. Niebur, K. W. Larsen, F. P. Seelos IV, N. O. Snider, and B. L. Joliff (2002), FIDO prototype Mars rover field trials, Black Rock Summit, Nevada, as test of the ability of robotic mobility systems to conduct field science, *J. Geophys. Res.*, *107*(E11), 8002, doi:10.1029/2000JE001464.

Backes, P. G., J. S. Norris, M. W. Powell, M. A. Vonn, R. Steinke, and J. Wick (2003), The science activity planner for the Mars Exploration Rover Mission: FIDO field test results, paper presented at Aerospace Conference, Inst. of Electr. and Electron. Eng., Big Sky, Mont., 8 –15 March.

Baker, V. R. (1999), Geosemiosis, *GSA Bull*, *111*(5), 633–645.

Bares, J. E., and D. S. Wettergreen (1997), Lessons from the development and deployment of Dante II, paper presented at Field and Service Robotics Conference, Canberra, Australia.

Brodaric, B., and M. Gahegan (2001), Learning geoscience categories in situ; Implications for geographic knowledge representation, paper presented at 9th International Symposium on Advances in Geographic Information Systems, Assoc. for Comput. Mach., Atlanta, Ga.

Cabrol, N. A., et al. (2001a), Nomad Rover Field Experiment, Atacama Desert, Chile: 2. Identification of paleolife evidence using a robotic vehicle—Lessons and recommendations for a Mars sample return mission, *J. Geophys. Res.*, *106*(E4), 7807–7815.

Cabrol, N. A., et al. (2001b), Nomad Rover Field Experiment, Atacama Desert, Chile: 1. Science results overview, *J. Geophys. Res.*, *106*(E4), 7785–7806.

Cabrol, N. A., et al. (2007), Life in the Atacama: Searching for life with rovers (science overview), *J. Geophys. Res.*, doi:10.1029/2006JG000298, in press.

Christian, D., D. Wettergreen, M. Bualat, K. Schwehr, D. Tucker, and E. Zbinden (1997), Field experiments with the Ames Marsokhod rover, paper presented at International Conference on Field and Service Robotics, Canberra, Australia, 7 –10 Dec.

De Hon, R. A., N. G. Barlow, M. K. Reagan, E. A. Bettis III, C. T. Foster Jr., V. C. Gulick, L. S. Crumpler, J. C. Aubele, M. G. Chapman, and K. L. Tanaka (2001), Observation of the geology and geomorphology of the 1999 Marsokhod test site, *J. Geophys. Res.*, *106*(E4), 7665–7682.

Ericcson, K. A., and H. A. Simon (1984), *Protocol Analysis: Verbal Reports as Data*, MIT Press, Cambridge, Mass.

Ferguson, R. S. (1991), Detection and classification of muskox habitat on Banks Island, Northwest Territories, Canada, using landsat thematic mapper data, *Arctic*, *44*, suppl. 1, 66-74, ASTIS record 31329.

Fink, W., J. M. Dohm, M. A. Tarbell, T. M. Hare, and V. R. Baker (2005), Next-generation robotic planetary reconnaissance missions: A paradigm shift; planetary and space, *Science*, *53*, 1419–1426.

Fong, T., H. Pangels, D. Wettergreen, E. Nygren, B. Hine, P. Hontalas, and C. Fedor (1995), Operator interfaces and network-based participation for Dante II, paper presented at 25th International Conference on Environmental Systems, Soc. of Automotive Eng., San Diego, Calif.

Franklin, S. E. (1991), Topographic data and satellite spectral response in subarctic high-relief terrain analysis, *Arctic*, *44*(1), 15–20.

Frodeman, R. (1995), Geological reasoning: Geology as an interpretive and historical science, *GSA Bull.*, *107*(8), 960–968.

Hine, B. P, et al. (1994), The application of telepresence and virtual reality to subsea exploration, paper presented at 2nd Workshop on Mobile Robots for Subsea Environments, Remotely Oper. Vehicles Comm., Monterey, Calif.

Latour, B. (1995), The "Pedofil" of Boa Vista," *Common Knowledge*, *4*(1), 87.

Lynch, M., and S. Woolgar (Eds.) (1994), *Representation in Scientific Practice*, MIT Press, Cambridge, Mass.

Matthews, S. B. (1991), An assessment of bison habitat in the Mills/Mink lakes area, Northwest Territories, using landsat thematic mapper data, *Arctic*, *44*, suppl. 1, 75-80, ASTIS record 31,330.

Nakamoto, J. (2006), Discrepancy investigation protocol: Applications to field and vicarious science, Masters thesis, Univ. of Iowa.

O'Connor, P. D. T. (2002), *Practical Reliability Engineering*, 4th ed., John Wiley, New York.

Pudenz, E. (2006), Directed autonomy: A new model for a new mobile robot technology, Masters thesis, Univ. of Iowa.

Pudenz, E., J. Glasgow, G. Thomas, P. Coppin, D. Wettergreen, and N. Cabrol (2006), Searching for a quantitative proxy for rover science effectiveness, paper presented at Conference on Human-Robot Interaction, Assoc. for Comput. Mach., Salt Lake City, Utah, 2–4 March.

Rasmussen, J. (1983), Skills, rules, and knowledge; signals, signs, and symbols, and other distinctions in human performance models, *IEEE Trans. Syst. Man Cybernetics*, *13*, 257–266.

Schumm, S. (1991), *To Interpret the Earth: Ten Ways to Be Wrong*, Cambridge Univ. Press, Cambridge, UK.

Seelos, F. P., IV, R. E. Arvidson, S. W. Squyres, E. T. Baumgartner, P. S. Schenker, B. L. Jolliff, C. S. Niebur, K. W. Larsen, and N. O. Snider (2001), FIDO prototype Mars rover field trials, May 2000, Black Rock Summit, Nevada, *Proc. Conf. Lunar Planet. Sci.*, *XXXII*.

Stoker, C., and B. Hine (1996), Telepresence control of mobile robots: Kilauea Mars robots: Kilauea Marsokhod experiment, Am. Inst. of Aeronaut. and Astronaut., Reno, Nev.

Thomas, G., M. Reagan, A. Bettis, N. Cabrol, and A. Rathe (2001), Analysis of science team activities during the 1999 Marsokhod Rover Field Experiment: Implications for automated planetary surface exploration, *J. Geophys. Res.*, *106*(E4), 7775–7784.

Thomas, G., J. Wagner, Z. Xiang, A. Kanduri, and J. Glasgow (2004), Analytical rover operations development, paper presented at Conference on Systems, Man, and Cybernetics, Inst. of Electr. and Electron. Eng., The Hague, Netherlands.

Tunstel, E., et al. (2002), FIDO rover field trials as rehearsal for the NASA 2003 Mars Exploration Rovers Mission, paper presented at Ninth International Symposium on Robotics With Applications, World Autom. Congr., Orlando, Fla.

Tunstel, E., T. Huntsberger, and E. Baumgartner (2004), Earth-based rover field testing for exploration missions on Mars, *Proc. World Autom. Congr.*, *15*, 307–312.

Vincoli, J. W. (1994), *Basic Guide to Accident Investigation and Loss Control*, John Wiley, New York.

Volpe, R. (1999), Mars rover navigation results using Sun sensor heading determination, paper presented at International Conference on Intelligent Robots and Systems, Inst. of Electr. and Electron. Eng., Kyongju, Korea.

Wagner, J., G. Thomas, J. Glasgow, R. C. Anderson, N. Cabrol, and E. Grin (2004), Error-associated behaviors and error rates for robotic geology, paper presented at Human Factors and Ergonomics Society 48th Annual Meeting, New Orleans, La., 20–24 Sept.

Walters, J. M., and R. L. Sumwalt III (2000), *Aircraft Accident Analysis: Final Reports*, McGraw-Hill, New York.

Warren-Rhodes, K., et al. (2007), Robotic ecological mapping: Habitats and the search for life in the Atacama Desert, *J. Geophys. Res.*, doi:10.1029/2006JG000301, in press.

Wheat, A. G. (2005), *Accident Investigation Training Manual*, Thomson Delmar Learning, New York.

Whittaker, W. R., D. Bapna, M. Maimone, and E. Rollins (1997), The Atacama Desert trek: A planetary analog field experiment, paper presented at International Symposium on Artificial Intelligence, Robotics and Automation in Space (i-SAIRAS'97), Tokyo, Japan, July.

J. Bretthauer, J. Glasgow, J. Nakamoto, I. U. Peate, E. Pudenz, and G. W. Thomas, Department of Mechanical and Industrial Engineering, University of Iowa, Iowa City, IA 52242, USA.

N. Cabrol, E. Grin, and K. Warren-Rhodes, NASA Ames Research Center, Moffett Field, CA 94035, USA.

C. Cockell, British Antarctic Survey, Cambridge CB3 0ET, UK.

P. Coppin, Eventscope, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

G. C. Diaz, Universidad Católica del Norte, 0610 Antofagasta, Chile.

J. M. Dohm, Department of Hydrology and Water Resources, University of Arizona, Tucson, AZ 85721, USA.

G. Fisher, N. Minkley, A. S. Waggoner, and S. Weinstein, Molecular Biosensor and Imaging Center, Carnegie Mellon University, Pittsburgh, PA 15213, USA.

A. N. Hock, Department of Earth and Space Sciences, University of California, Los Angeles, CA 90095, USA.

L. Marinangeli and G. G. Ori, IRSPS, 65127 Pescara, Italy.

J. Moersch and J. L. Piatek, Department of Earth and Planetary Sciences, University of Tennessee, Knoxville, TN 37996, USA.

T. Smith, K. Stubb, M. Wagner, and D. Wettergreen, Robotics Institute, Carnegie Mellon University, Pittsburgh, PA 15213, USA.