What Changes When a Large Team Becomes More Expert? Analyses of Speedup in the Mars Exploration Rovers Science Planning Process

Irene V. Tollinger
(irene.tollinger@nasa.gov)
NASA Ames Research Center
Moffett Field, CA 94035, USA

Christian D. Schunn
(schunn@pitt.edu)
LRDC, University of Pittsburgh
Pittsburgh, PA 15260, USA

Alonso H. Vera
(alonso.vera@nasa.gov)
Carnegie Mellon & NASA Ames
Moffett Field, CA 94035, USA

Abstract

We present data from the Mars Exploration Rovers (MER) '03 mission to unpack the factors that produce a speed-up of group performance in science planning. Data was collected from a wide variety of sources via systematic sampling (including video, ethnographic, and *in situ* "tick" coding) as well as complete coverage (including daily presentations, reports, and rover plan files) The work examines a broad set of hypotheses which range from selection of easier plans to increased reuse. Though all the hypotheses were plausible, five of the ten were found to have support including more work done individually and decreased engineering uncertainty. The goal is to better understand the factors that affect group performance and to make predictions in applied contexts.

Introduction

With increasing levels of practice, time on task drops and gradually approaches an asymptote. This speed up is a hallmark of expertise (Anderson, 1982; Anderson, 1993; Logan, 1988; Newell & Rosenbloom, 1981; Newell & Simon, 1972). Expertise has been studied in depth at the individual level (Ericsson & Smith, 1991; Chi, Glaser, & Farr, 1988). At the group level, expertise has also been studied (Hutchins, 1995) but is often measured by proxies such as productivity (Argote, 1999). We are interested in understanding the particular factors that affect a speed up of group performance. To this end, we present data from the Mars Exploration Rovers (MER) '03 mission. The analyses to be presented will enhance understanding of the relevant factors and improve predictions of group performance.

The paper has the following structure. The introduction covers the mission domain and the speed up phenomenon itself. The methods section covers data collection techniques and is followed by the results section which covers each hypothesis tested and associated findings in detail. The discussion section provides a synthesis and considers the implications of the findings for design of collaborative tools.

Mission Domain

The purpose of the MER mission is to further Mars exploration through the deployment of twin robotic rovers outfitted with a payload of scientific instruments. Launched in June of 2003, the rovers landed in January 2004 and nominal surface operations proceeded through April 2004 successfully completing the 90 sols (Martian days) of operations. Both rovers have continued to perform well and are in an extended phase of operations as of the writing of this paper (January 2006). The rovers, referred to as MER A and MER B, can perform several hours of activity per sol (e.g. taking photographs or driving) contingent on the availability of limited resources such as battery power.

The MER Science Team, known as the Science Operations Work Group (SOWG) will be the focus of analysis in this paper. They provided science staffing for both rovers. For each daily science shift, the team was organized into five groups by discipline: Geology, Mineralogy, Rock & Soil, Atmospheres, and Long Term Planning (LTP). Groups were characterized by semi-fluid membership. A scientist generally worked in a particular group but could move back and forth (e.g. working a 3 to 4 day stint in Geology, taking a few days leave, spending a stint in LTP, and returning to Geology). Groups usually consisted of between two and eight scientists on a given day. The entire team was led by the SOWG Chair who worked to guide the team to ask the right questions, explore better alternatives, and build consensus. The SOWG Chair also had final decision authority. Team structure and facilities were identical for both rover teams.

Each day the science team, collocated in the Science Assessment Room at the Jet Propulsion Laboratory, was tasked with generating a set of science activities that would be processed by the engineering team and sent to the rover to be executed the following day. This involved: understanding the hard constraints for the day (e.g. power levels, data volume available for transfer), forming a sense of the type of day it would be (primarily devoted to taking a large panorama versus deploying an array of sensing instruments on a particular rock surface versus driving), assessing which activities the rover successfully completed the previous day, developing each science activity request (e.g., take a 30 second image of the sky at x, y, z location), and negotiating the priority of these requests among the whole team. The process was reliant on specific hardware and software tools; some of the tools were developed to support collaboration, but most were not and yet all were used collaboratively on occasion.

Because the rovers are solar powered, mission planners decided to run nominal operations on Mars time so that they could consistently send the rover a new plan each Martian morning. This meant that the scientists and engineers synchronized their activities to the local time on Mars for their rover (called Local Solar Time or LST). Mission planners originally allotted 8.5 hours to science planning for each sol. The science schedule included the following elements (times reflect the original schedule):

Science Context meeting: 11:00-11:30 LST

- Determine which activities were or were not executed the previous sol
- Discuss/decide the type of sol to plan for tomorrow (drive, panorama, etc.)

Science Downlink Assessment meeting: 16:00-17:00 LST

- Discuss instrument and rover health
- Assign activities to particular science theme groups to plan

SOWG meeting: 18:00-20:00 LST

- Refine and prioritize the requested activities in light of resource limitations

Between the Science Context meeting and the Science Downlink Assessment meeting (4.5 hours), the team was provided time to look at new data products as they came in, discuss the options for the following sol, and develop activity requests using the Science Activity Planner application. The hour between the Science Downlink Assessment meeting and the SOWG meeting was provided primarily to develop/complete activity requests.

The ~15.5 remaining hours, post science planning, were allotted to engineering work necessary to schedule the science requests according to resource constraints, convert the requests into sequences the rover could execute, and perform verification and validation. The schedule was identical for both rovers though shifted by 12 hours as their landing sites were on opposite sides of the planet. The overall planning cycle was considered tight. Analogous missions had used operations concepts with a 2-3 day planning cycle to generate a single day's worth of rover commands but such a structure would have cut down drastically on the amount of science gathered.

Phenomenon

Unexpectedly, a "speed up" phenomenon emerged. By mission day 85, the science planning process on both missions had been cut from 8.5 hours to approximately 2.5 (MER A) and 2 hours (MER B). The graph in Figure 1 represents the change in the length of the science shift (from the beginning of the Science Context meeting to the end of the SOWG meeting) over time. The decisions to shorten the science shift, seen as discrete steps on the graph, represent decisions made by relatively high-level mission people as temporary changes to the schedule that were then made permanent. The data comes from official daily mission schedules.

After the end of the nominal mission (post sol 90), the mission entered an "extended" operations phase with the allotted time set to 2 hours for science planning for both missions. The extreme brevity of the schedule in this phase suggests that the science schedule, decided before the beginning of the mission, was incredibly compressible after 90 days of practice (MER A r = .8, MER B r= .97).

The speed up phenomenon was also evaluated in regards to the length of the daily science meetings. Meeting length was coded from video data for both the Science Context meeting and Science Downlink Assessment meeting. The Science Downlink Assessment meeting shows a decrease (MER A r= -.56, MER B r= -.48) while the Science Context meeting (which sets the science agenda for the day at the most general levels of planning) shows little change (MER A r= -.24, MER B r= -.36). Perhaps planning is more compressible than science or perhaps the more detailed group planning is more compressible than very general planning. Overall, though there was some compression of meetings, it seems that much of the speed up was due to cuts in the 5.5 hours of work time allotted between meetings.

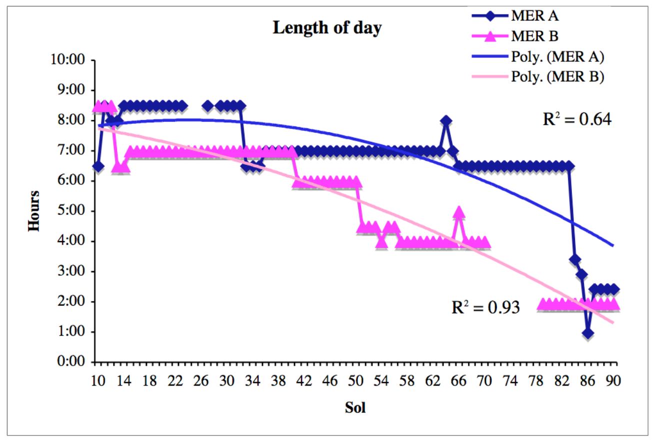

Figure 1. Length of science day as a function of time on MER A and B.

It is valuable to explore and understand the factors that contributed to this change in order to both better predict performance of future missions and assess the nature of the speed up (e.g. less science being done versus the same amount of science being done faster) to potentially foster and better support this type of speed up.

Methods

Data collection covers the first 90 sols of the nominal mission for each rover, via a number of data sources. For the observational data, we used a counter-balanced sampling design. Researchers collectively made 20 trips. Each researcher made one early and one late trip on each of the two rovers. A trip consisted of 3 days (~8 hours per day). The total was ~100 hours of observation per person. Observational data collected include: (1) video data of science theme group areas, (2) audio data collected at workstations, (3) in-situ tick sheet coding (# people doing specific task/ tool) described below, and (4) ethnographic data. In addition to the data collected in person, the authors have mined a variety of digital data sources including: mission staffing schedules (excel, database), documentarian notes (MSWord), activity plan files (Rover Markup Language), meeting presentation files (PPT), online reports by engineering/science role (HTML document). Together, these data are used to test the hypotheses presented below.

The *in situ* tick sheet coding was intended to quantitatively measure what the scientists were doing in the science theme groups during the work periods between the daily meetings. Each minute, for fifteen minutes, a researcher would code the number of people/tasks on paper "tick" sheets. For example, two people are independently using Word on a laptop and two people are jointly using the Science Activity Planner within a science theme group.

During the 5.5 hours of work time between meetings (given the original 8.5 hour day), a researcher would code 30 minutes per science theme group in a systematic sampling sequence (2.5 hours of coding). As the number of hours per day was cut down, the sampling had to be reduced. In the end, the work yielded an N of 4053 minutes (67.6 hours) in which a researcher observed activity and an N of 8340 ticks.

The tick sheet coding process was pilot-tested during a 10-day operational readiness test (Field Integrated Development and Operations) the year before the mission began. The contents of the tick sheets were validated from videotape and a high level of coding reliability was found.

Our statistical analysis approach is as follows. For data collected for all 90 sols, we compute a mean measure per sol and then look for linear correlations between sol and that measure. For data collected less often, we created blocks of time (e.g., early, middle, late) and then calculated the mean measure for each block of time. All statistical results reported as significant conform to the $p < .05$ level.

Results

We present a wide-ranging analysis, in terms of both data sources and hypotheses. The goal is as much to rule out potential contributing causes as to identify the root causes of the speed up phenomenon. The hypotheses described below were developed based on the deep ethnographic knowledge of the context gathered in the course of the series of operational readiness tests leading up to the mission and the mission data collection itself. The set of hypotheses evaluated are presented in four groups. The first two groups encompass hypotheses for which we did not find support, where the observed factors stayed consistent over the duration of the missions (or went in the opposite direction). Each of these are further divided into those that are essentially confounds (they explain the speedup by assuming the task itself was changed to a simpler task, and no real speedup occurred) and those that were considered core potential explanations of a real speedup in planning. The second two groups cover hypotheses for which we did find support and are also split into confounds and core explanations.

Potential Confounds Ruled Out

Hypothesis 1: The number of people decreased

Important but rare and risky events tend to have many extra people around in the beginning, and then only the core, necessary group remains as the hours accumulate. If there were fewer people participating in science planning, this would change the planning task itself and could cause a decrease in planning time simply due to fewer options being proposed each day. While the lead science roles were officially staffed by the mission (SOWG chair, LTP lead, and LTP documentarian), the majority of science theme group members followed a less formal schedule. Science team members are generally busy researchers with external academic affiliations who traveled to the Jet Propulsion Laboratory in southern CA from their home institutions in the US and abroad. *In situ* tick sheet data were queried for the average number of people in the theme groups per sol across both missions. Variability in the data due to people temporarily flowing across groups should not pose analysis problems given that it should be similar for all groups and there is a large N (4053 minutes coded).

The results show that overall the number of scientists did not decrease significantly over time (MERA $r= .34$, MER B $r= .12$). Based on this data it appears unlikely that the planning task was changing due to a decrease in staffing levels.

Hypothesis 2: Amount of science planned decreased

It is possible that the science team planned fewer science activities on average per sol in response to the decreases in the amount of time allotted. The 'speed up' would then be somewhat spurious, representing a scenario in which planning speed is consistent but less science is accomplished. The MER mission organized planning requests into a two-tier hierarchy. Activities are the lower level construct. They represent the instrument to be used and all the necessary parameters (e.g., use the Panoramic camera to take an image at X, Y, Z location with the red filter). Individual activities were also grouped into observations, a higher-level scientific construct, based on a shared scientific goal (e.g. all the images of the really interesting rock we called Adirondack).

The results showed that the number of observations increased significantly for MER A but not for MER B (MER A $r= .29$, MER B $r= .18$). The number of activities follows the same pattern with a significant rise for MER A but not for MER B (MER A $r= .36$, MER B $r= .14$) Based on this data, it is unlikely that the cause of the decrease in science planning time was associated with a decrease in the amount of science planned. Given that the amount of science increased, the speed up phenomena is actually greater than the simple compression of the daily schedule.

Potential Core Explanations Ruled Out

Hypothesis 3: Easier plans were selected

A number of problem solving theories have posited that people are rational decision-makers, preferring efficient choices over inefficient choices, where efficiency includes the time spent planning and thinking and the probability of success (e.g., Anderson, 1990; Gigerenzer & Todd, 1999; Oaksford & Chater, 1998; Siegler, 1996). In a complex domain like MER, it will take time for the problem solvers to determine which kinds of activities involved significantly more planning time. Over time, it is possible that task speedup occurs as the problem solvers begin favoring activities that involve significantly less planning time.

In advance of the mission, the science team developed a formal set of designations for the type of plan the rover would execute at a high level. There are five such "sol types" (in rough order from easiest to most difficult): Panorama, Spectroscopy, Scratch and Sniff, Drive, and Trench. Sol type was tracked in an official decision tree document. While there is some variability within each sol type (e.g. some drives are longer and more complex than others), this is a reasonable approximation of complexity at the level of the individual plan/sol.

For this analysis, the 90-sol mission is split into blocks of 10 sols each. There was little correlation between any of the sol types and time represented as blocks—the largest significant correlation was r= .25 for Scratch and Sniff on MER A. As Scratch and Sniff is a relatively difficult sol type, a growth rather than decrease in its frequency argues strongly against this hypothesis. In fact, there is consistent variability in sol type choice over the course of the mission. Based on this analysis, it is unlikely that changes in sol difficulty influenced the efficiency of the science planning process.

Hypothesis 4: Longer term planning

In general, planning is done to improve problem solving accuracy and efficiency (Hayes-Roth, 1979). In a similar way, longer term planning can simplify more immediate planning. More specifically in the MER context, longer-term plans could lead to reduced time spent in the discussion of trade offs. For example, scientists consider the possibility of not taking a certain reading tomorrow and deciding to drive the following day. In such a scenario, they'd miss a certain type of data useful in comparing similar rocks. The knowledge of the most likely plan and its alternatives well into the future could support more efficient planning (e.g. the alternatives for the following day actually were panorama or trenching so rover will be at the same location to take the desired image).

The MER mission kept an official branching decision tree document called a "sol tree". It was based on the high-level goal for each sol. The measure used is the number of days projected ahead in the daily sol tree document. The results show a significant increase. On MER A, the growth was after sol 60 (r= .44). On MER B, growth was more cyclical (r= .27). Because the pattern of growth was inconsistent or primarily at the end of the mission, it is unlikely that this factor played a strong role in the speedup.

Hypothesis 5: Reuse of old plans increased

A potential method for decreasing the amount of time spent on science planning is to reuse plans from past sols. In many domains, such as software development, people often find a similar piece of work they've done, perform a "save as" operation, and modify the original work as appropriate to save time. In the MER context, this would involve the use plans from previous sols. Indeed, reuse of old plans is perhaps the most basic model of expertise: people retrieving prior solutions (Logan, 1988; Siegler & Shipley, 1995).

The MER mission kept all plans in a large server-based file system such that all previous plans were accessible to members of the science team. However, the act of opening an older plan and copy/pasting in a piece of that plan (observation or activity) or copying/pasting within the same plan was not logged. Therefore, the authors used an indirect measure. Associated with each observation was a set of free-text "intent" fields. These four fields, Notes, Purpose, Method, and Scientific Hypothesis, were there for scientists to fill out with a few words to a paragraph of explanation so that their intent would be available for members of the engineering team to reference. After the science shift is over, engineers may have to make modifications based on resource limitations. Thus, it is in the scientists' interest to convey what they are attempting to accomplish so that the data gathered is not compromised by uninformed modifications. The measure chosen is the provenance of intent fields. Provenance is categorized as: reuse (copied from an old plan), new (never seen in any previous plan, empty (no content), or duplicate (copied from within the plan of the current sol).

Here the results were somewhat mixed. On MER A, reuse increases significantly (r= .25). The different fields show different trends (Notes r= .09, Purpose r= .39, Method r= .09, Scientific Hypothesis r= -.3). On MER B, reuse seems to actually go down (Notes r= -.2, Purpose r= -.2, Method r= -.1, Scientific Hypothesis r=-.2). Though there is variability in the data between missions, the intent field data does not provide consistent evidence that the gains in science planning efficiency are due to reuse of plans from previous sols.

Potential Confounds With Support

Hypothesis 6: The amount of documentation decreased

The MER mission process as designed, required the science and engineering teams to document their work products, rationales for requests, etc. As the science team became more comfortable with their work, it is likely that the amount of documentation decreased. This is almost a "human nature" argument in that processes are followed as designed for a period but after a certain point people start to cut down on work that is not strictly critical path such as documentation. The measure here comes from analysis of intent field sources. The four intent fields, Notes, Method, Purpose, and Scientific Hypotheses exist to capture the science rationale and for members of the engineering team to reference later in the planning process.

The results show that the incidence of empty fields increases significantly for both missions (MER A r= .3, MER B r= .34). The results show a more consistent increase for MER B (Notes r= .46, Method r= .56, Scientific Hypothesis r= .32) than for MER A (Notes r= .03, Method r= .04, Scientific Hypothesis r= .31) where there is some variability among the fields. Based on this evidence, the amount of documentation does decrease over time though the decrease is not equally consistent for both missions.

Hypothesis 7: Available resources (power) decrease

A systematic decrease in resources can affect the planning process because fewer activities could be planned each sol, which would take less time. Each sol the mission tracked the projected power cost of the plan developed, the power remaining from the previous day, and the projected power accumulation from the solar panels.

The results show a strong correlation. Over time, the amount of available power decreased for both missions by approximately 35% in a fairly regular and gradual fashion (MER A = .99, MER B = .91). Available power decreased over the course of the mission as the seasons changed and it became winter on Mars (shorter days, lower temperatures, etc.). By itself this factor is somewhat inconclusive. Though there was less power over time and this may have contributed to the speed up in planning, the number of activities planned did not decrease over time. One possible interpretation could be that the number of activities did not change but with less power each activity became simpler (e.g. fewer filters to specify per image). It is important to note, however, that while power was an important constraint, it was only one of many constraints, and many sols involved plans that did not fully use the available power. For example, the correlation between sol and power consumed was lower (MER A r=.36, MER B r=.56).

Potential Core Explanations With Support

Hypothesis 8: More planning work was done individually or in small groups

If more planning work was done individually, the speed up of group performance is then simply a story of improved individual performance. Alternatively, the groups could have discovered that it was more efficient to work collaboratively in smaller sub-groups over time, and smaller subgroups may function more efficiently.

The *in situ* tick sheet coding provided data on the number of people doing a task and the tool they were using as described in the methods section. That data was split into early, middle, and late periods of the mission.

The overall results do not show an appreciable increase in individual work. Further analysis was performed on collaboration around the primary technologies including personal laptops, workstations, and a category we termed "no technology" (printed images, notebooks, talking etc.). The workstations housed the primary science data analysis and planning software (Science Activity Planner). The data shows that collaboration on workstations decreased over time (see Figure 2).

The results also show that group size remained relatively stable between the early and late periods of the mission. Of all collaboration observed, over 75% was small group collaboration (groups of 2 and 3) so there was little room for subgroups to form. Groups of 4 and 5 account for about 20% of the overall collaboration observed and groups of 6 or more account for 5% or less. This same pattern held for workstation use as well. Thus, the change is in the frequency of collaboration not the form of collaboration.

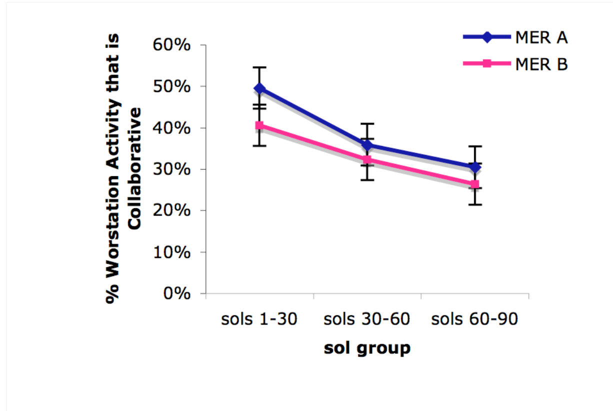

Figure 2. Percentage of work done collaboratively on workstations across early, middle and late time periods.

Hypothesis 9: Less engineering uncertainty

A potential factor in the efficiency of planning is the science team's knowledge of the rover capabilities and resources. In the context of MER, uncertainty is high early on in the mission. Rover design is based on best approximations for variables like landing site terrain, quantity of airborne dust as it influences available solar power, temperature, etc. As these initial values and their daily variance become better understood, the science team and their engineering team representatives would spend less time debating the safety and likelihood of success for particular activity requests.

Indicators of engineering uncertainty were coded for in the transcripts of video data collected during work periods between formal meetings. These variables were coded blind to sol and by research assistants unaware of the hypotheses under evaluation. All data was double-coded and inter-rater reliability was over 90%. We define uncertainty as not being known, fixed, or completely certain. First, utterances were classified as being "related to data analysis" versus "related to planning" or off-task. Second, all the on-task utterances were coded for psychological uncertainty (filtering out common types of verbal filler phrases such as "uh/um", "I mean", "you know"). Possible subtypes of uncertainty included qualifying, hedging, or estimating statements.

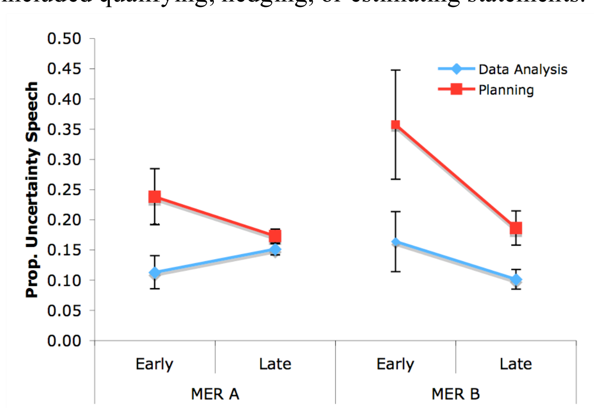

Figure 3. Proportion of uncertainty speech during planning and data analysis activities, early and late.

The data is split into early and late time periods. The results, shown in Figure 3, reveal that for planning related speech, uncertainty shows a consistent drop on both missions. For science data analysis related speech, uncertainty seems to be less consistent and the slight change is negative for one mission and positive for the other. Given that the end product of the science process is a plan for the following sol, it seems that the data support the effect of decreased engineering uncertainty on planning.

Hypothesis 10: Familiar roles

Practice within a given role should improve individual performance over time. The mission decided ahead of time that leadership positions would be staffed by a small set of people. Individual scientists within science theme groups did not have role-distinguished positions. The measure of role familiarity for each sol is the number of previous sols of experience accumulated by the person filling a given role. Data was taken from official staffing schedules, from roles listed in documentarian notes, and from personnel listed on daily science meeting presentations. The analysis is performed on the role of SOWG chair, the leader and final decision-maker for the science team.

The results show a correlation between role familiarity and time for both missions (MER A r= .78, MER B r=.67). The familiar roles hypothesis seems to be supported. This suggests two alternative interpretations. It is possible that improving group efficiency over time is a consequence of gradually improving leadership. Alternatively, it is possible that, like leaders, each individual gradually improves in his/her position, leading to overall better group performance. Both of these effects may play a role.

Discussion

The hypotheses tested above are a subset of the overall hypotheses originally developed. In future work, we will analyze the remaining hypotheses including: stronger interpersonal relationships, more formalized decision-making and science discussions, larger granularity of planning, and greater willingness to compromise and wait.

Ten hypotheses were examined here, reflecting the complex nature of group work. A priori, it was possible that all ten hypotheses would have received some support in our correlational analyses, producing little real gain in our scientific understanding of group expertise. In fact, only five factors had evidence of change, and these factors distributed themselves across different levels of explanation. The mission itself changed somewhat, in that power resources decreased consistently over time. Psychologically, the mission also changed in that there was less engineering uncertainty about what the rovers could do. Individuals simplified their own tasks by engaging in less documentation. Some core work became less collaborative. And finally, leaders became more familiar with their leadership roles. It is likely that most if not all of these factors would only allow for a moderate speedup, as they themselves changed only a moderate amount over the mission. Thus, it is probable that their combination was required to produce the sizeable speedup that was observed.

The above analyses also highlight the value of applying combined ethnographic and quantitative methods. Without the ethnographic observation, it would have been difficult to generate hypotheses as to the factors that affect the science planning process. The value of the quantitative data was in the surprising results that emerged. It is interesting that the factors that proved relevant (less documentation, fewer (power) resources, more work done individually, less uncertainty, and familiar roles) are not the more individual-expertise factors that adaptive software usually supports, such as moving common actions to the top of menus. The application of adaptive interface technology to address changes in individual expertise would not work here—people didn't just settle on certain actions, or begin to reuse old plans with high frequency via copy/paste actions. We have a challenge to propose what new paradigms of adaptive software would support changing familiarity with leadership roles and changes in uncertainty levels.

Acknowledgments

NASA's Science Mission Directorate (SMD) supported this work.

References

Anderson, J. R. (1982). The acquisition of cognitive skill. *Psychological Review*, *89*, 369-403.

Anderson, J. R. (1990). *The Adaptive Character of Thought*. Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1993). Problem solving and learning. *American Psychologist, 48*, 35-44.

Argote, L. (1999). Organizational Learning: creating, retaining, and transferring knowledge. Kluwer Academic Publishers.

Chi M, Glaser R, & Farr M, (eds.) (1988). *The Nature of Expertise*. Hillsdale, NJ: Lawrence Erlbaum.

Ericsson A, & Smith J. eds. (1991). *Towards a General Theory of Expertise: Prospects and Limits*. Cambridge, U.K: Cambridge University Press.

Gigerenzer, G. & Todd, P. M. (1999). Fast and frugal heuristics: The adaptive toolbox. In Gigerenzer & Todd/ABC Group (Eds.), *Simple Heuristics that Make Us Smart* (pp. 3–34).

Hayes-Roth, B. & Hayes-Roth, F. (1979). A cognitive model of planning. *Cognitive Science, 3*, 275-310.

Hutchins E. (1995). How a cockpit remembers its speeds. *Cognitive Science, 19*, 265-288.

Logan, G. D. (1988). Toward an instance theory of automatization, *Psychological Review, 95*(4), 492-527.

Newell, A. & Rosenbloom, P. (1981). Mechanisms of skill acquisition and the law of practice. In J. R. Anderson (Ed.), *Cognitive skills and their acquisition* (pp. 1-55). Hillsdale, NJ: Erlbaum.

Newell, A. & Simon, H. A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Oaksford, M. & Chater, N. (1998). *Rationality In An Uncertain World: Essays on the Cognitive Science of Human Reasoning*. East Essex, UK: Psychology Press Ltd.

Siegler, R. S. (1996). *Emerging minds: The process of change in children's thinking*. New York: Oxford University Press.

Siegler, R. S., & Shipley, C. (1995). Variation, selection, and cognitive change. In G. Halford & T. Simon (Eds.), *Developing cognitive competence: New approaches to process modeling* (pp. 31-76). New York: Academic Press.