



Unobtrusive Monitoring of Spaceflight Team Functioning

Literature Review and Operational Assessment for NASA Behavioral Health and Performance Element

*Veronica Maidel, M.S.
School of Information Studies
Syracuse University, Syracuse, NY*

*Jeffrey M. Stanton, Ph.D.
School of Information Studies
Syracuse University, Syracuse, NY*

*Kathryn E. Keeton, Ph.D.
NASA Johnson Space Center, Houston, TX*

THE NASA STI PROGRAM OFFICE . . . IN PROFILE

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA Scientific and Technical Information (STI) Program Office plays a key part in helping NASA maintain this important role.

The NASA STI Program Office is operated by Langley Research Center, the lead center for NASA's scientific and technical information. The NASA STI Program Office provides access to the NASA STI Database, the largest collection of aeronautical and space science STI in the world. The Program Office is also NASA's institutional mechanism for disseminating the results of its research and development activities. These results are published by NASA in the NASA STI Report Series, which includes the following report types:

- **TECHNICAL PUBLICATION.** Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA's counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.
- **TECHNICAL MEMORANDUM.** Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.
- **CONTRACTOR REPORT.** Scientific and technical findings by NASA-sponsored contractors and grantees.

- **CONFERENCE PUBLICATION.** Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or cosponsored by NASA.
- **SPECIAL PUBLICATION.** Scientific, technical, or historical information from NASA programs, projects, and mission, often concerned with subjects having substantial public interest.
- **TECHNICAL TRANSLATION.** English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services that complement the STI Program Office's diverse offerings include creating custom thesauri, building customized databases, organizing and publishing research results . . . even providing videos.

For more information about the NASA STI Program Office, see the following:

- Access the NASA STI Program Home Page at <http://www.sti.nasa.gov>
- E-mail your question via the internet to help@sti.nasa.gov
- Fax your question to the NASA Access Help Desk at (301) 621-0134
- Telephone the NASA Access Help Desk at (301) 621-0390
- Write to:
NASA Access Help Desk
NASA Center for AeroSpace Information
7115 Standard
Hanover, MD 21076-1320



Unobtrusive Monitoring of Spaceflight Team Functioning

Literature Review and Operational Assessment for NASA Behavioral Health and Performance Element

*Veronica Maidel, M.S.
School of Information Studies
Syracuse University, Syracuse, NY*

*Jeffrey M. Stanton, Ph.D.
School of Information Studies
Syracuse University, Syracuse, NY*

*Kathryn E. Keeton, Ph.D.
NASA Johnson Space Center, Houston, TX*

Available from:

NASA Center for AeroSpace Information
7115 Standard Drive
Hanover, MD 21076-1320
301-621-0390

National Technical Information Service
5285 Port Royal Road
Springfield, VA 22161
703-605-6000

This report is also available in electronic form at <http://ston.jsc.nasa.gov/collections/TRS/>

1. TABLE OF CONTENTS

2. Executive Summary	1
3. Introduction	2
4. Industrial Performance Monitoring	4
Reactions to Performance Monitoring	5
Effects and Outcomes of Performance Monitoring	6
Empirical Research on Effects and Outcomes	6
Conceptual models of Performance Monitoring	7
5. Brief Orientation to Team Dynamics and Performance	8
6. Overview of Possible Indicators	9
7. Mental Models	10
Mental Models and Team Mental Models - Definitions	10
Linking Team Mental Models and Team Outputs	12
Methods to extract and measure team mental models	14
Cognitive interviewing techniques	15
Verbal protocol analysis	15
Visual card sorting technique	15
Ordered tree technique	16
Causal mapping	16
Content analysis	17
Observation of task performance	17
Pathfinder	18
Multidimensional scaling	18
Cognitive mapping Techniques	18
Team Mental Models: Conclusion	22
8. Extracting team and individual characteristics from text	23
Text Analysis of Team Member Discourse	23
Extraction of Emotions from Text	25
9. Biometric Methods	28
10. Proxemics	30
11. Overall Conclusion of The Literature Review	32
12. Products Overview	33
Introduction	33
Collecting Textual Communication	36
Text Analysis Packages	40
13. A Case Study of <i>PolyAnalyst</i> Software	46

14. A Case Study of <i>LIWC</i> Software	49
15. Summary of Interviews with Key Personnel	53
16. Overall Conclusions	55
17. Future Steps and Research Recommendations	56
18. References	59

ACRONYMS

ANEW	Affective Norms for English Words
AT	Automatic Tagging
CD	Communications Density
EPM	Electronic Performance Monitoring
HLT	Human Language Technologies
IECM	interactively elicited causal map
LC	Lag Coherence
LEW	List of Emotional Words
LSA	Latent Semantic Analysis
MOS	Multidimensional Scaling
NLP	Natural Language Processing
PF	Pathfinder
TBCM	Text-based Causal Maps
UAV	Unmanned Aerial Vehicle

2. EXECUTIVE SUMMARY

This document contains a literature review suggesting that research on industrial performance monitoring has limited value in assessing, understanding, and predicting team functioning in the context of space flight missions. The review indicates that a more relevant area of research explores the effectiveness of teams and how team effectiveness may be predicted through the elicitation of individual and team mental models. Note that the “mental models” referred to in this literature typically reflect a shared operational understanding of a mission setting such as the cockpit controls and navigational indicators on a flight deck. In principle, however, mental models also exist pertaining to the status of interpersonal relations on a team, collective beliefs about leadership, success in coordination, and other aspects of team behavior and cognition.

Pursuing this idea, the second part of this document provides an overview of available off-the-shelf products that might assist in extraction of mental models and elicitation of emotions based on an analysis of communicative texts among mission personnel. The search for text analysis software or tools revealed no available tools to enable extraction of mental models automatically, relying only on collected communication text. Nonetheless, using existing software to analyze how a team is functioning may be relevant for selection or training, when human experts are immediately available to analyze and act on the findings. Alternatively, if output can be sent to the ground periodically and analyzed by experts on the ground, then these software packages might be employed during missions as well. A demonstration of two text analysis software applications is presented.

Another possibility explored in this document is the option of collecting biometric and proxemics measures such as keystroke dynamics and interpersonal distance in order to expose various individual or dyadic states that may be indicators or predictors of certain elements of team functioning. This document summarizes interviews conducted with personnel currently involved in observing or monitoring astronauts or who are in charge of technology that allows communication and monitoring. The objective of these interviews was to elicit their perspectives on monitoring team performance during long-duration missions and the feasibility of potential automatic non-obtrusive monitoring systems.

Finally, in the last section, the report describes several priority areas for research that can help transform team mental models, biometrics, and/or proxemics into workable systems for unobtrusive monitoring of space flight team effectiveness.

Conclusions from this work suggest that unobtrusive monitoring of space flight personnel is likely to be a valuable future tool for assessing team functioning, but that several research gaps must be filled before prototype systems can be developed for this purpose.

3. INTRODUCTION

In the context of space flight, teamwork is an essential ingredient in successful missions. A variety of adverse influences may negatively impact the performance of mission teams both on the ground and in flight. Such influences may include physical stressors on the organism such as diurnal disruption, effects of microgravity, injury, or task overload as well as psychological factors such as social isolation, role overload, or interpersonal conflict among team members. Given the importance of team effectiveness, NASA's Behavioral Health and Performance Element (BHP) has identified a need to monitor the functioning of teams, primarily using unobtrusive means. The purpose of such monitoring lies in providing a stream of indicators that can serve several operational goals:

1. Monitoring during personnel selection activities can provide input for the selection of compatible team members and of individuals with psychological profiles suited to teamwork in extreme environments and situations.
2. Monitoring during training activities can provide diagnostic information useful in guiding further instruction and coaching as well as in determining the composition of teams prior to mission deployment.
3. Monitoring during missions can provide forewarning of potential operational failures due to disruptions of team functioning and give the opportunity to take preventative measures.

These purposes of monitoring make sense only if the collected indicators, whether gathered unobtrusively, through self report, or by other means, are reasonably predictive of outcomes of interest. These outcomes may include subjective and objective assessments of team task performance, team safety performance, accidents, and team-level psychosocial outcomes such as cohesion and morale. In psychometric terms, all indicators obtained from monitoring must be valid assessments of team functioning and must be predictive of some mission outcome of interest.

Unobtrusive monitoring techniques are preferable in the scenarios described above because they would not require the active involvement of personnel in provision of the measures. In addition, given that teams will work in a variety of remote environments, it can be assumed that technology-mediated methods of capturing behavior and communications will be required, because direct observation by supervisors, coaches, or psychologists will generally be feasible only during selection and training activities. With that being said, self-report measures and other assessments that require the active participation of team members may be valuable during a validation phase.

The literature review in this document provides an overview of prior research on the various methods of monitoring personnel performance and the effects that monitoring

have on job performance and on other outcomes. Most of this research has arisen from industrial contexts and may not have universal relevance to the space flight context. As a result, we have expanded our view of the literature to include consideration of some areas that have typically not been considered in the research realm of performance monitoring, but may yet provide some worthwhile insights.

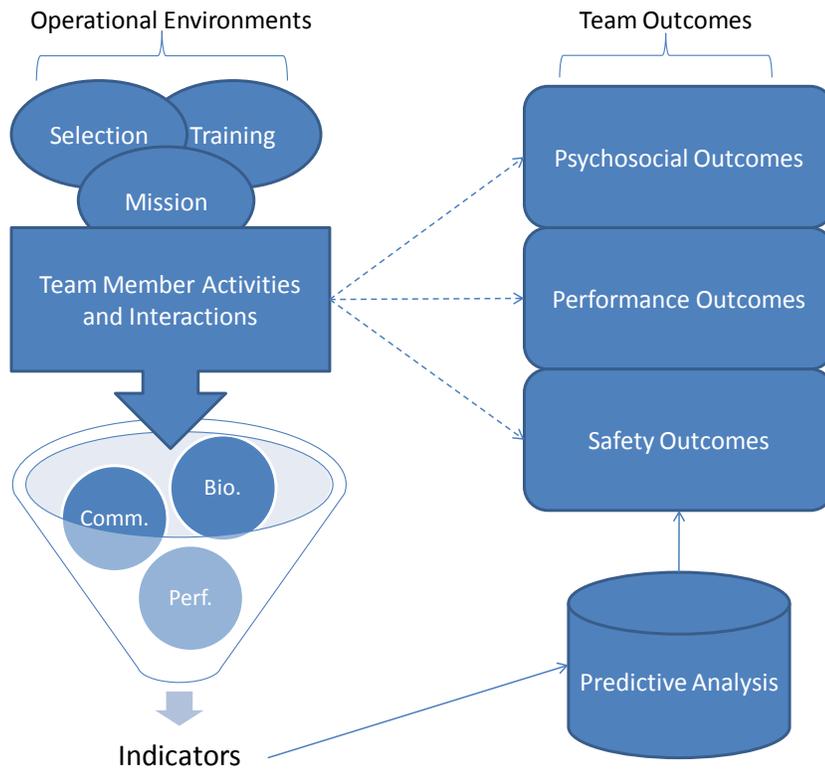


Figure 1: Conceptual overview of the problem space.

Throughout our analysis of the literature conducted to date, we have assumed that the ultimate goal of this project is to assess the feasibility and options for solutions that would combine unobtrusive collection of indicators with predictive analysis. This assumption is embodied in Figure 1 which reflects our understanding of the problem space. Starting at the top left of Figure 1, we have imagined three operational contexts: a selection context where individual team members are chosen for various roles in a mission; a training context where team members may work together on simulated or practice tasks; and a mission context where the team functions during space flight or in other mission environments. In the latter two contexts, team members interact, perform individual tasks, and collaborate on group tasks. These activities presumably cause the various outcomes experienced during training or missions (three dotted lines pointing right).

We have gathered these outcomes into three gross categories: psychosocial outcomes (e.g., morale, cohesion); performance outcomes (e.g., task completion); and safety outcomes

(e.g., mistakes, accidents). We expect that many behaviors and activities reveal observable cues or indicators about the functioning of a team (the funnel shape on the left).

Communicative indicators may include speech and textual communications among team members and between team members and those on the ground. Performance indicators may include intermediate task results and work products (e.g., completion of a subtask in a repair job), physical interactions among team members (e.g., assisting another team member with equipment), or timing indicators (e.g., sleep-wake schedules, time-on task). Finally, another set of indicators focuses on biometrics such as infrared detection of surface blood flow, urinalysis, and galvanic skin response.

To close our consideration of Figure 1, we assume that among the various unobtrusive indicators of individual and team activity, a subset of such indicators may have predictive value in foreshadowing important outcomes such as changes to morale, team performance, or the occurrence of accidents. The cylinder at the lower right of Figure 1 represents an analysis component in which indicators are combined, scored, normed, and compared in an effort to predict outcomes of interest. Throughout the literature review below, we have, in effect, “graded” the existing research with respect to whether we believe it provides promise with respect to indicators, analysis, and/or prediction.

4. INDUSTRIAL PERFORMANCE MONITORING

To provide a context for the various indicators that could predict team outcomes, we will start by looking at traditional performance monitoring as it has been conducted in industrial settings. In industrial environments such as call centers and manufacturing floors, performance monitoring refers to the gathering of indicators about the work effectiveness and productivity of individuals, groups, and larger organizational units. Prior to the widespread deployment of information and communications technologies, supervisors monitored performance by personally observing, recording, and reporting on employee behavior and work products (Attewell, 1987; Stanton & Julian, 2002). Technological advances over the past 40 to 50 years, such as inexpensive personal computers and networks, have facilitated new techniques for performance monitoring and encouraged widespread deployment of these techniques.

Psychologists, sociologists, and others have raised concerns about the use and effects of performance monitoring in the industrial workplace (Stanton & Julian, 2002). In non-military work environments, workers have a range of legal rights – variable across different countries – that influence when, where, and how performance monitoring technologies may be used. Additionally, in many industrial contexts, the existence of labor markets means that work conditions are a source of competitive advantage. As a result of the labor market effects and/or the possibility of employee litigation, many researchers have focused their efforts on understanding how employees *react* to the use of

performance monitoring in their work environments. Reactions to performance monitoring have raised sufficient concern that researchers have developed specialized self-report scales specifically for this purpose, such as the one published by Flint (2008). The next section provides a brief overview of this research area.

REACTIONS TO PERFORMANCE MONITORING

Research on reactions to performance monitoring identifies and explores employees' attitudes and perceptions and relates these concepts to subsequent outcomes. Dependent variables in this research include job attitudes (including fairness, job satisfaction, organizational commitment, and obligation to reciprocate) (Wells, Moorman, & Werner, 2007; Watson, 2008), stress (D. Kiker & M. Kiker, 2008), and mood state (Davidson & Henderson, 2000). For example, the meta-analysis by Kiker and Kiker (2008) showed that electronic performance monitoring was negatively correlated with employee job satisfaction and positively associated with job stress. Researchers have also examined contingent factors that influence reactions and attitudes, such as the existence of feedback (Alder, 2007; Alder & Ambrose, 2005), personality and demographic attributes (J. V. Chen & Ross, 2007), organizational cultures (Alder, 2001), and prior beliefs (Alder, Schminke, Noel, & Kuenzi, 2008).

The issues considered by researchers of industrial monitoring tend to have the greatest relevance in a setting where employees are not used to being monitored, where there are issues of employee retention, and where employees are represented by unions. The space flight context is substantially different in several ways. For example, space flight personnel are monitored frequently on their physical health, are highly familiar with the purposes and goals of self-report measures, and have substantial commitment to activities with demonstrable connections to mission success or safety.

Consistent with Alder (2001), who argues that more "bureaucratic" organizational cultures will respond more favorably to monitoring than supportive cultures, it is reasonable to expect that highly trained space flight personnel may not have the same reactions to performance monitoring as is observed among workers in an industrial environment. Rather than a concern for basic labor rights, space flight personnel may have concerns for the time or inconvenience of monitoring techniques. Space flight personnel also may have concerns around personal privacy, particularly given the confined size of their operational environments.

These concerns suggest that while space flight personnel may have *different* reactions than, say, call center workers, it is no less important to have the cooperation and "buy-in" of space flight personnel with respect to deployed monitoring techniques. An important lesson from the use of monitoring in industrial environments is that employees are creative in finding ways to circumvent controls and mechanisms they consider objectionable

(Stanton & Stam, 2006). Therefore, when designing and implementing any monitoring tool for the space flight environment, it will be essential to involve space flight personnel in the processes of evaluating and deploying monitoring tools. Emphasizing to the personnel the advantages of performance monitoring (e.g., safety) and assuring them that they will be protected from the consequences of revealing their mistakes is important as well. Additionally, it is important to be aware of other effects and outcomes that performance monitoring may inadvertently influence. Such effects are discussed in the following section.

EFFECTS AND OUTCOMES OF PERFORMANCE MONITORING

Another question addressed by numerous studies is how performance monitoring affects job/task performance and other related outcomes. Performance refers either to the individual performance within a team or the performance of the whole team, which could be measured with quality of output and quantity of output. Most of the experimental studies in this vein focus on relatively simple clerical tasks such as sorting and editing. Other outcomes refer to safety, errors, and psychosocial aspects that may be applicable to a team, such as team cohesion and morale. Some of the literature in this area focuses on “surveillance,” which is a subtype of monitoring used to uncover wrongdoing (D'Urso, 2006). In the industrial context, such monitoring may help to ensure that employees are not stealing, performing sabotage, or procrastinating.

EMPIRICAL RESEARCH ON EFFECTS AND OUTCOMES

OUTPUT QUALITY AND QUANTITY

Quality and quantity of task output are of interest in the present project and therefore it is important to review what previous empirical research has shown when exploring how performance monitoring affects output quality and output quantity. Many studies have focused on how task difficulty influences the quality and the quantity of output given that a performance monitoring system is present.

Davidson & Henderson (2000) found that participants performing an easy task displayed increased task performance under electronic performance monitoring (EPM) and poorer performance when performing a difficult task under EPM. Similarly, Park & Catrambone (2007) sought to investigate whether “virtual humans” embodying the role of the performance monitoring system produced social facilitation effects. They found that for easy tasks, performance in a “virtual human” monitoring condition was better than in the alone condition, and for difficult tasks, performance in the “virtual human” condition was worse than in the alone condition. Consistent with these results from individual studies, in a meta-analysis of EPM literature, Kiker and Kiker (2008) found that EPM has a positive effect on performance *quantity* but a negative effect on performance *quality*. They also ascertained that the EPM-performance quality relationship was moderated by task difficulty such that EPM improved performance quality for simple tasks, but detracted from

it for complex tasks. Social facilitation, a theoretical perspective that considers neural system activation and arousal as a basis for changes in performance, is frequently harnessed to explain these effects.

Working in a different theoretical vein, Stanton & Julian (2002) concluded that workers' perceptions of importance of a task were influenced by the capabilities of electronic performance monitoring even though, in all cases, a supervisor stated that both quality and quantity of performance were important. Workers perceived quality to be more important when quality was the only aspect of the task that the system monitored. Workers perceived quality performance to be of lesser importance when only the quantity of performance was monitored. These results suggest that the psychological effects of monitoring with respect to focusing attention and motivating behavior through expectations can be an unintended side effect of both the design of a monitoring system and the communications that managers use to explain and justify the system.

FEEDBACK

Some of the research in industrial monitoring focuses on feedback provided to the employees through monitoring systems. Although performance monitoring has typically been construed as a supervisory activity, the data that monitoring produces can just as easily be used in feedback processes with workers. This line of research generally does not examine the quality-quantity trade-off but rather takes a general view of performance improvement. For example, Alder (2007) found that allowing employees to determine when they receive feedback may enhance their desire to improve. In turn, to the extent that perceptions of interpersonal fairness are high, individuals' desire to respond to feedback will result in improved performance. Similarly, Goomas (2007) discovered that immediate performance feedback and self-monitoring that was delivered to employees improved order picking performance. This improvement was due to an intervention package that included the depiction of goal times and immediate performance feedback.

CONCEPTUAL MODELS OF PERFORMANCE MONITORING

The performance monitoring literature contains conceptual models and frameworks guided by psychological theories such as the theory of planned behavior, social facilitation, and the theory of procedural justice. These models portray the relationships between the various factors and outcomes of performance monitoring.

For example, Moran & Nakata (2009) proposed a model based on the Theory of Planned Behavior (Ajzen, 1991) that examines adverse effects caused by ubiquitous monitoring. The theory of planned behavior holds that specific attitudes toward a behavior can predict the occurrence of that behavior. In addition to measuring attitudes toward the behavior, people's subjective norms (their beliefs about how people they care about will view the behavior in question) are also measured (Ajzen, 1991). The factors in Moran & Nakata's

model influence factors from the Theory of Planned Behavior and include context, justification (which is affected by trust), awareness, control, boundaries, and intrusion. Behavioral intentions in the Theory of Planned Behavior eventually affect the two outcomes: intended work behavior and unintended work behavior.

Other effects of electronic monitoring are explored by D'Urso (2006) who examined the "panoptic" effects (i.e., a fear of continuous surveillance) of monitoring interpersonal communication in the workplace. According to the model developed in this study, outcomes such as organizational fairness, job performance, workplace satisfaction and others are influenced by organizational management style, organizational communication climate, comfort with technology, and surveillance beliefs.

Cultural dimensions of monitoring were investigated by Panina & Aiello (2005) who proposed a model describing the interaction of major EPM characteristics and national culture dimensions, and suggesting possible implications of this interaction on creating culture-sensitive EPM designs.

The papers reviewed above demonstrate the range of concerns and variables emphasized in performance monitoring research in recent years. It is evident that the focus has been mostly on industrial environments, where the outcomes of interest and the factors influencing these outcomes (such as organizational management style, organizational communication climate, job attitude, fairness perceptions) reflect common characteristics of industrial labor markets. Workers who belong to unions, or who are willing and able to quit a position, or who can raise legal challenges to adverse working conditions have influenced researchers' decisions about which contexts, variables, and organizations to examine. On a related note, the industrial use of electronic performance monitoring has occurred most frequently in environments where managers are concerned that unmonitored workers may exhibit unproductive or counterproductive behaviors. As a result, the workers and tasks that are monitored tend to be relatively unskilled.

To conclude, although the empirical and conceptual literature on the effects of monitoring is quite thick, much of the literature is only indirectly applicable to the space flight environment. In the literature, performance monitoring is rarely applied to teams and rarely used in the context of highly technical or high level professional jobs. We make a set of baseline assumptions about space flight personnel concerning their levels of organizational commitment, motivation, and task performance that suggest we must look elsewhere in the research for ideas about monitoring the status and functioning of teams.

5. BRIEF ORIENTATION TO TEAM DYNAMICS AND PERFORMANCE

Although team dynamics and performance are familiar topics to many readers of this material, we provide a brief overview of them here to uncover a few assumptions that are

important to the remainder of this paper. A team is generally defined as a small group of interacting individuals charged with performance of a task, set of tasks, or mission (Guzzo, 1995). The group's membership is well bounded, and members identify with the group.

Research on teams indicates that team performance is a "cross-level" construct, dependent on both the capabilities and characteristics of individual team members and the quality of interaction among them. Interactions among team members fit into a modest number of functional categories: coordination is an important example of a communicative activity that helps keep a team functioning effectively. Leader directives, conflict management, and goal-setting represent other common areas of communication.

These communication processes lead a team through various developmental stages, during which differentiated social and performance roles emerge among team members (Hare, 2003). For example, even in teams without formally assigned leadership, one or more leaders tend to emerge over time. Effective differentiation of roles together with effective enactment of those roles positively influences individual satisfaction, team morale, team cohesion, and team performance.

6. OVERVIEW OF POSSIBLE INDICATORS

From the brief discussion of teams above, it is evident that indicators of team functioning can be obtained both from individual team members and from interactions among team members. Communications among team members that reveal common understandings of roles, tasks, and goals (as well as areas of dissensus) can provide an important window into how well a team is functioning. Biometric indications of anger and other physiological/emotional states that occur during team interactions, or stress states that persist following team interactions may also provide useful indications of team functioning. Finally, team outcomes, such as intermediate task completion or time-on-task, when compared with established norms or benchmarks, may provide indirect evidence of the quality of team dynamics.

As suggested by Figure 1, a combination of several indicators may help to predict psychosocial, performance, and safety outcomes. In terms of communicative indicators, team mental models, which reflect a shared operational understanding of a mission setting, as well as the status of interpersonal relations on a team, collective beliefs about leadership, and other aspects of team behavior and cognition, may be elicited from textual communications through textual analysis. Along with team mental models, extraction of emotions from text in order to represent the general state of mind of the team and individuals is also a viable option.

Biometric indicators such as keystroke dynamics, facial expressions, gestures, speech, skin temperature, galvanic skin response, and electromyography (muscle activity) may provide

another source of information to complete the full picture of how a team of astronauts functions during selection, training or mission. This document will focus mainly on emotion identification at a single moment in time, but a more appropriate usage of a biometric system might track and attempt to identify patterns of change in emotions over time.

Finally, physical interactions among team members (e.g., communicative nonverbal indicators), may be assessed using a strategy known as “proxemics,” an area of research that focuses on the perception, use, and structuring of space. When dealing with proxemics, most often researchers study how spatial use affects and reflects relationships between individuals as members of a dyad or a larger group, and whether the particular use of space is intentional (i.e. seeking interaction) or inadvertent (i.e. in a public setting). In the space flight context, one challenge would be to take advantage of movement and body position in three dimensions. A second challenge lies in the automatic identification and coding of proxemic measures.

Interestingly, some researchers who have been interested in the performance of teams, have approached it from the perspective of underlying cognitive mechanisms instead of overt behavior or motivation. As Rouse et al. (1992) suggested, deficiencies in team coordination, communication and overall performance may be better understood by focusing on underlying mechanisms rather than global behaviors.

One may conceive that the tools for discovering and exposing these underlying mechanisms are a form of performance monitoring, but one that focuses on precursors of complex team activities rather than directly upon the activities themselves. Some researchers who have examined these precursors have focused on “mental models” of complex task performance held by individuals and teams. Research on mental models provides an opportunity to understand how to collect information about a team that may predict later team performance on complex tasks. The next section of the review examines the mental models literature and the techniques used by researchers in this area to extract mental models.

7. MENTAL MODELS

MENTAL MODELS AND TEAM MENTAL MODELS - DEFINITIONS

As systems and technologies utilized in the workplace became more complex over the last 20 to 30 years, the issue of individual mental models started gaining interest among researchers. Research had shown that understanding a complex system (e.g., a cockpit) and successfully interacting with it required several different types of knowledge, including knowledge of the basic system components, the possible states of those components, and how the components are interrelated (Hegarty, 1991; Rowe & Cooke, 1995). Such

knowledge comprises a mental representation, or "mental model," of the system (Gentner & Stevens, 1983; Rowe & Cooke, 1995; Staggers & Norcio, 1993). A worker who operates complex equipment or system interfaces uses a mental model to understand the systems and any feedback that they provide (Rasmussen & Jensen, 1974; Rowe & Cooke, 1995). Although a well-articulated mental model is not always necessary for effective interactions with complex equipment, mental models are assumed to play an important role in facilitating most human-system interactions, particularly when the equipment behaves in an expected manner (Rowe & Cooke, 1995).

Complex systems often require several operators to work together in order to achieve a goal. One relevant example is the space shuttle's remote manipulator system, which typically requires two coordinated operators for safe and effective use. In such a scenario, each individual needs a well-developed response pattern to external events and the actions of other operators. Thus, shortly after mental models began triggering interest among researchers, "team mental models" also began to gain importance in these research communities. Klimosky and Mohammed's (1994) definition of a team mental model asserts that it is an emergent characteristic of the group that is more than just the sum of individual mental models. Although the measurement techniques used to capture team mental models are on the individual level, a team mental model is a group-level phenomenon. As described by the definition, team mental models are team members' shared, organized understanding and mental representation of knowledge or beliefs about key elements of the team's relevant environment.

According to Klimoski and Mohammed, team mental models reflect organized knowledge, internalized beliefs, assumptions, and perceptions. Usually it will be in the form of a set of concepts stored and retrieved from memory in relationship to one another. Such organization may derive from presumed cause and effect linkages, or it may reflect learned patterns. Moreover, while the organized patterns may be "spatial" or "sequential" in nature, most probably, such knowledge is organized semantically. The content of shared mental models might reference representations of tasks, of situations, of response patterns or of working relationships. Allowing for the impact of method and circumstances of measurement, a team mental model represents how the group members as a collectivity think or characterize a set of phenomena associated with effective team performance of complex tasks (Klimoski & Mohammed, 1994). It is possible that multiple mental models (or multiple facets of a single model) coexist simultaneously among team members at a given point in time. These would include models of task/technology, response routines, team work, and social relations (Klimoski & Mohammed, 1994).

With respect to the present review, team mental model research as currently represented in the literature probably has the greatest relevance when imagining a team performance monitoring solution that predicts complex task performance. In contrast, the review below

suggests that few if any efforts in the team mental model literature have pertained to the psychosocial status or outcomes of the team. Nonetheless, we present a quite thorough review of the area below in the belief that some of the unobtrusive mental model assessment techniques might eventually be harnessed in support of understanding a broad range of team performance criteria.

LINKING TEAM MENTAL MODELS AND TEAM OUTPUTS

Klimoski and Mohammed (1994) argued that the following are important for linking shared mental models with team performance: communication processes, strategy and coordinated use of resources, and interpersonal relations or cooperation. A particular team member must have a conceptualization of what is expected of him or her by each team member for each to jointly succeed (Klimoski & Mohammed, 1994). Team mental models are constructed of both the aggregation of individual mental models regarding the task and the technology, but also of the mental models of how a team operates and what role each team member needs to take.

A great deal of research has been directed by the assertion that since teamwork mental models guide the manner in which individuals perform their tasks and interact with one another, team members who hold *similar* or *aligned* mental models of teamwork are better able to coordinate with one another and thus achieve superior performance outcomes (Mathieu, Heffner, Goodwin, Salas, & Cannon-Bowers, 2000). It has been hypothesized that team mental models enable team members to form common expectations, coordinate actions, adapt their behaviors to task demands, facilitate information processing, provide support, and diagnose deficiencies. As such, team mental models influence both team processes (e.g., communication, conflict) and team outputs (Klimoski & Mohammed, 1994; Kraiger & Wenzel, 1997; Marks, Mathieu, & Zaccaro, 2001).

Several studies have been conducted to explore the relationship of team mental models with various team outputs. One of the first on this subject was Rouse et al. (1992) who considered the nature of team performance in complex systems in a context of military training of command and control. Rouse showed that the mental models construct has the potential to provide the basis for a principled explanation of team performance, as well as an avenue for enhancing performance. More specifically, Rouse argued that usage of the mental models construct in terms of the mechanisms underlying the formation of expectations and explanations may enable development of finer-grained understanding of such global team-related phenomena as coordination and communications performance.

Stout et al. (1999) explored the relationship between team planning, shared mental models, and coordinated team decision making and performance in surveillance/defense missions using a commercially available low-fidelity helicopter simulation. Results indicated that effective planning “increased” the shared mental model among team

members (indexed as the similarity of individual models to a collective model), allowed them to utilize efficient communication strategies during high-workload conditions, and resulted in improved coordinated team performance.

Mathieu et al. (2000) also used a flight simulator in their study to examine the influence of convergence, or sharedness, of team members' mental models as related to team processes and performance. The general results showed that team processes were related significantly to team performance. More detailed analyses in the same study revealed that team mental model sharedness related significantly to team performance, but the relationship was fully mediated by team processes (e.g., coordination of activities).

Edwards, Day, Arthur, & Bell (2006) used a video game that was designed to simulate a complex and dynamic aviation environment to examine the relationship between the similarity and accuracy of team mental models and compared the extent to which each predicted team performance. The authors presented evidence that, for a task with a defined set of optimal strategies, team mental model accuracy is a stronger predictor of team performance than team mental model similarity. In this case, accuracy was operationalized by comparing trainees' mental models to an expert referent model that served as the "true state of the world." Unlike other research that tends to favor similarity, this pattern of results did not emerge until later in training. In an attempt to explore the determinants of team mental models, this study also provided evidence that team members' ability is related to the development of similar and accurate mental models and that the accuracy of mental models partially mediates the relationship between team ability and team performance.

A study addressing similar constructs by Lim and Klein (2006) examined the relationship between team mental model similarity and accuracy and the performance of combat teams. The teams were expected to perform under high stress and intense time pressure. Their findings suggested that teams whose members organize and structure their team related knowledge in a similar fashion will find it easier to coordinate their activities. These team members are likely to agree on team priorities and strategies, yielding efficient task performance. Additional findings suggested that team mental model accuracy was also instrumental for team performance. Teams whose average mental models were most similar to experts' mental models performed better than did teams whose average mental models were less similar to experts' mental models.

Team mental model similarity was also explored by Smith-Jentsch et al. (2001), and their findings indicated that higher-ranking navy personnel held mental models of teamwork that were more similar to an empirically derived model of expert team performance, than lower-ranking personnel. Furthermore, comparisons of mental model similarity within groups of high- and low-ranking trainees and within groups of high- and low-experience

trainees indicated greater similarity between those of higher rank and between those with greater experience. Another study by the same authors tested the effects of a computer-based training strategy that was designed to develop teamwork mental models that were more similar to the “expert model” described in the previous study. Using a card sorting approach, positive training effects were demonstrated on similarity to the expert model, similarity to other trainees, and consistency.

As the review above suggests, individual and team mental models have been widely researched in work environments that may have similarities to those where space flight personnel train and work. Moreover, mental models are utilized when the team or individuals need to tackle complex tasks, which is also the more suitable case for our purposes. Finally, the strong (if complex) connection between team mental models and team performance suggests that results from this body of research may be quite relevant to the space flight context. In the next section, we describe the wide variety of possible mental model elicitation techniques. Although many of these techniques are obtrusive and suitable only for research studies, the literature does suggest some possibilities for operational contexts.

METHODS TO EXTRACT AND MEASURE TEAM MENTAL MODELS

The means to measure, elicit, or represent mental models in general and team mental models in particular have been discussed extensively in literature on personnel training. Incorporating mental model assessment, diagnosis, and instruction into training requires the selection of an appropriate measure of the knowledge, structure, and assertions in mental models. Because there is no universally agreed-upon measure of this knowledge, selection of a measure can be difficult (Rowe & Cooke, 1995). It has always been challenging to determine the best way to measure mental processes of organized knowledge because these processes are tacit, residing in the person’s mind. Therefore, the elicitation of mental models has been a central issue in individual and team mental models research, and various methods have been proposed to extract the information that represents mental processes. Some papers have been written specifically for this purpose while others elaborate on this issue in detail in the methods section due to its importance.

Langan-Fox, Code, & Langfield-Smith (2000) constructed a review describing the potential of each technique for individual and team mental model elicitation and representation. According to the authors, different elicitation techniques require different degrees of researcher involvement, and some techniques are more suited to eliciting an individual mental model than a team mental model. Following are some of the elicitation techniques presented in this review.

COGNITIVE INTERVIEWING TECHNIQUES

This category of techniques includes interviews, question-answer interviews, and a technique called inferential flow analysis. A transcript of the interview is constructed and analyzed using propositional or discourse analysis. The final representation is a graph that illustrates domain concepts along with conditional and causal associations among them. Cognitive interviewing techniques can be used to elicit a team mental model directly through group discussion. These group discussions can be used to derive important constructs within a domain and linkages and relationships between those constructs. A disadvantage of group discussion is that, like in any group discussion, often the views of more influential or extraverted group members can dominate the discussion and distort the team mental model in favor of their perspectives. This can be partially overcome by asking each individual to write down his or her responses before group consensus is achieved. An example of application of such technique was an investigation of changes in managers' mental models through the extensive review of questionnaires, interviews and company records (Cavaleri & Sterman, 1997).

VERBAL PROTOCOL ANALYSIS

This technique is used primarily to obtain information about decision-making strategies and general reasoning processes. It is particularly useful for uncovering decision-making errors attributable to individual biases and misconceptions. Participants are asked to think aloud while they undertake a task or make a decision. Sessions are recorded on audiotape or videotape, and a written protocol is generated afterward. From the set of recorded verbalizations, the researcher can identify the relationships between objects within a domain. Possible outputs from this technique include sets of production rules, decision trees, heuristics, algorithms, systematic grammar networks, and more. A disadvantage of this technique is that the individual-level output produced by verbal protocol analysis might be difficult to summarize and compare systematically, which limits the usefulness of the technique for team mental model measurement. This technique was applied in the examination of thinking processes in personnel selection (Barber & Roehling, 1993) as well as for physicians' medical reasoning and problem solving (Hassebrock & Prietula, 1992).

VISUAL CARD SORTING TECHNIQUE

This technique is a quick, easy-to-administer, flexible, and face-valid way of representing mental models. In visual card sorting, the participant is either provided with researcher-generated concepts or is asked to list all the concepts that he or she sees as relevant to the domain of interest. The concepts are written on cards, and the participant is asked to sort the cards by placing cards that are perceived to be related closer together. The participant then explains why he or she arranged the cards in such a way. This information is tape recorded or transcribed, and the arrangement of cards (the final representation) is photographed. Although the visual card sorting technique can be used in a group session to

measure the team mental model, as with cognitive interviewing techniques, the views of more influential or extraverted group members can dominate the session and distort the model. The use of visual card sorting for team mental model measurement is recommended when research time is limited. One of the studies that used this technique (Daniels, de Chernatony, & Johnson, 1995) examined managers' mental models of competitive industry structures.

ORDERED TREE TECHNIQUE

This technique was created as an alternative to multidimensional scaling, when researchers observed that the recall of items in a free recall task often included consistent sequencing of items recalled. While multidimensional scaling suggests that items recalled together may have a "short" distance from each other, the ordered tree technique also considers consistencies in the sequence of items recalled (e.g., when recalling "a," one tends to next recall "b," but not vice versa). In this technique, participants are asked to recall a large, well-learned set of items many times from many different starting points, sometimes starting with a cue item and sometimes without. An algorithmic analysis constructs a hierarchical structure among the items based on the resulting sequences. The basic assumption is that respondents have mentally organized items into chunks and will recall the chunks as units, tending to recall a whole chunk before proceeding to the next one. An example of ordered tree technique usage was the investigation of the long-term effects of teacher education programs on beginning teachers' cognitive structures for classroom management (Winitzky, Kauchak, & Kelly, 1994).

The ordered tree technique can be used to compare hierarchies between pairs of respondents. Measures of similarity can be calculated between a pair of trees. Perhaps more importantly, team members can discuss the similarities and differences between the hierarchies as a training exercise. The method has often been applied to research that focuses on mental model similarity in expert-novice comparisons.

CAUSAL MAPPING

In this technique, the participant is asked whether one concept influences the other, whether it does so positively or negatively, and if it does so weakly, moderately, or strongly for each possible pair of a set of concepts. An $n \times n$ adjacency matrix is then constructed where n is the total number of concepts in the map, and numbers in the cells at the intersection of each column and row indicate the existence, direction, and strength of the relationship between two concepts. A distance ratio formula can be used to infer the extent of difference between the maps of individual team members. An example of such usage of a distance ratio formula is Langfield-Smith (1992) who investigated the collective beliefs about the important aspects of the job of a fire protection officer in a team of firefighters. Markíczy & Goldberg (1995) inspected causal mapping for an individual and proposed a

method for expanding causal mapping's value as a tool for exploring individual's idiosyncratic beliefs.

The techniques described above require the presence of a researcher to conduct and guide the process of elicitation. Many of the techniques are time intensive and, in some cases, research participants find them annoying. Alternative techniques described below are less obtrusive in the sense that they work from the analysis of incidentally produced materials (such as formal speeches or the recording of the team or individual in action).

CONTENT ANALYSIS

This is a family of systematic methods for analyzing written statements such as formal speeches and transcripts of interviews. The researcher uses a set of coding rules to analyze sentences, phrase by phrase, to uncover important concepts and the relationships between them. Establishing the validity of content analysis for deriving a team mental model is problematic. Content analysis is applied to a corpus of textual data that can be obtained from a variety of sources, such as emails or reports. The purposes of these communications, the circumstances under which they were produced, and the intended audience all influence the information available for analysis. Under optimal circumstances, when a corpus of communications is explicitly focused on the coordination of a team's tasks, it may be possible to derive useful information about individual and team mental models. As a more obtrusive method, interview transcripts can also be used as the corpus for content analysis. One example is the work of Langan-Fox & Tan (1997) who applied content analysis on interview transcripts to investigate organizational culture within a large, government business enterprise.

OBSERVATION OF TASK PERFORMANCE

Researchers can use direct observation of an individual's behavior during the completion of a task to infer mental models. Although complete observation involves a high level of involvement between a researcher and participant, passive observations are also possible. Passive observation entails little or no interaction between the two parties, and the researcher often takes the role of a bystander or uses technological means such as activity logs or videos of task activity to provide indirect evidence of a mental model. A difficulty with passive observation is that it is up to the researcher to identify the important concepts and the relationships between them, and behavior is not always a good guide. For example, a mistake in controlling a system may be due to an inaccurate mental model or simply due to inattention or fatigue. Observation of task performance is best suited to the examination of (individual) mental models in contexts where a user must interact extensively with a system and the sequence of interactions, mistakes, backtracking, and related actions illustrate the nature of the individual's mental model. An example of such work is by Chen (1996) who looked at students' interpersonal cognitive problem-solving skills where behavior observation was one of the methods to elicit those skills. Vandenplas-Holper

(1996) used video recordings of children's learning sessions and analyzed using systematic observation to infer changes in mental models over time with increased learning.

Mohammed, Klimoski, & Rentsch (2000) also evaluated a set of techniques for measuring team mental models: Pathfinder, multidimensional scaling, interactively elicited cognitive mapping, and text-based cognitive mapping were critiqued and compared according to their treatment of content and structure, as well as their psychometric properties.

PATHFINDER

Pathfinder (PF) is intended to produce psychological scaling of the underlying structure between concepts. The PF algorithm transforms raw, paired comparison data into a network structure in which the concepts are represented as nodes and the relatedness of concepts is represented as links between nodes. Studies that use PF employ an “averaging” technique to transform individual-level data into a team-level cognitive structure. The team level structure can then be compared back to individual structures, and members could then be asked to verify the map for accuracy. It is also feasible for group members to work jointly in order to rate the similarity between constructs and produce a team-level map.

MULTIDIMENSIONAL SCALING

Multidimensional scaling (MDS) is a psychometric scaling technique that represents proximity data in a spatial map. Given the assumption that geometric distance can represent psychological similarity, MDS can be useful in identifying the unknown underlying dimensions used to cognitively organize stimuli. MDS represents cognitive structures in n-dimensional space. Inputs are most commonly in the form of similarity ratings that respondents provide for pairs of items. The resulting MDS solution, calculated based on similarity data, presents stimuli in relation to the underlying dimensions. Studies that use MDS also average individual-level data to examine team-level cognitive structures. As with PF, no known examples of global measurement exist. However, group members could jointly rate the similarity between constructs to produce a team-level map.

COGNITIVE MAPPING TECHNIQUES

Cognitive mapping methods are graphic representations of both the content and structure of individuals' personal belief systems in a particular domain. Cognitive mapping was one of the first cognitive measurement techniques to be introduced into management research and has been used to study decision making, negotiation, organizational cognition, and strategy. Cognitive mapping provides a way of accessing large, untapped sources of data generated by organizations and examines meaning as a relational phenomenon. There are two techniques by which the content to be mapped can be generated. The first, called an interactively elicited causal map (IECM) is obtained by requesting the data from participants through questionnaires and/or interviews and the second, text-based causal

map (TBCM), is obtained through post hoc analyses of data (e.g., systematic coding of documents or transcripts).

A similar approach called *map analysis* was employed by Carley (1997), who used this method for extracting, analyzing, and combining representations of individual mental models as cognitive maps. This textual analysis technique allows the researcher to extract cognitive maps, locate similarities across maps, and combine maps to generate a team map. Using map analysis, the researcher can address questions about the nature of team mental models and the extent to which sharing is necessary for effective teamwork. Individual cognitive maps can be compared, or combined to create a team cognitive map. If two individual's mental models have been coded as cognitive maps, then these maps can be compared and contrasted. Each individual cognitive map can be thought of as a binary graph (an acyclic, usually tree-shaped structure). As such, they can be compared simply by counting the number of shared concepts, shared statements, total concepts, total statements, concepts only in that map, and statements only in that map. Two maps can be combined by creating either a union or intersection file.

Carley (1997) demonstrated this technique using data drawn from a study of software engineering teams. The impact of critical content analysis coding choices on the resultant findings was examined. Various coding choices were shown to have systematic effects on the complexity of the coded maps and their similarity. Consequently, a thorough analysis requires analyzing the data several times under different coding choices. A substantive result reported by Carley is that all teams have comparable models, but successful teams are able to describe their models in more ways than are non-successful teams.

Rowe & Cooke (1995) conducted an empirical study that evaluated four measures to assess individual mental models, with individual task performance as the criterion. The authors tested three methods that involved technicians who tried to deal with a troubleshooting problem: a laddering interview, relatedness ratings, diagramming, and think aloud. Some of these methods are similar to the methods described by Langan-Fox et al. (2000) and Mohammed et al. (2000), but this study also attempted to test these methods' ability to predict performance.

In a *laddering interview*, after being given a troubleshooting problem statement, the technician was asked: (1) to identify the major system important in troubleshooting this problem; (2) to name the major components of the identified system in the context of the troubleshooting problem; and (3) to list all the major components of the identified system, regardless of the problem's context.

For *relatedness ratings*, the technician used a six-point scale to rate the functional relatedness of all pairs of the 11 system components. Pairs were presented randomly, and

technicians were told to rate them in terms of their first impression of functional relatedness, within the context of the troubleshooting problem.

In the *diagramming task*, the technician arranged and connected index cards, with a component name printed on each, in a manner that represented the way in which the system functions in general. Connections and their directionality were represented with a set of directional and bidirectional arrows.

For the *think aloud* task, the technician stated the troubleshooting actions he or she would take, and a subject matter expert stated the results of those actions. The technician was instructed to verbally express all thoughts, or think aloud, while working to solve the problem.

Of the four techniques assessed, all but the think-aloud technique were predictive of troubleshooting performance. Although the think-aloud verbal reports yielded mental models, these models were not predictive of performance. This may have been because this technique is unstructured, and structuring the think-aloud interview might have resulted in data more closely related to performance. This finding emphasized the importance of verifying that the mental model measure relates to the criterion of interest. The laddering and rating techniques were independently predictive of performance, suggesting that these two measures capture different aspects of a mental model, each of which is important to the troubleshooting task. The laddering task tapped into knowledge about existing components, whereas the ratings task accessed knowledge about the interfaces or connections between components. Both of these measures appeared to be good choices for identifying mental model knowledge when the goal is to improve troubleshooting performance.

While the previous study focused mainly on individual mental models, Smith-Jentsch et al. (2001) reported results from two empirical studies that utilized a card sorting approach to measuring team member mental model similarity in naturalistic training environments. The authors adopted an expert model of teamwork that was derived through the analysis of performance ratings collected from navy command and control teams. This model consisted of four dimensions defined by 11 component behaviors: information exchange (i.e., passing information, providing big picture summaries, seeking information from all available sources); communication (i.e., proper phraseology, brevity, clarity, completeness of standard reports); supporting behavior (i.e., error correction, back-up/assistance); and leadership (i.e., providing guidance, stating priorities).

A card sorting task was used to assess each participant's mental model of teamwork. Each card listed a concrete example of either effective or ineffective teamwork that could occur in a submarine attack center. Participants were instructed to sort the examples into categories of teamwork that were meaningful to them and to label each of their piles.

Participants' similarity to one another and to the expert model was computed based on matrices ("1" was placed in each cell where the corresponding cards were placed together in a single category) by using the Phi coefficient (the Pearson correlation coefficient between two dichotomous variables). To obtain an expert matrix from which to score the accuracy of participants' mental models, three researchers sorted the examples into piles that would be consistent with the expert model of teamwork.

The elicitation strategy presented in Smith-Jentsch et al. (2001) differed from many of those mentioned above in other articles in that most of the team mental model research aspired to extract an aggregated mental model for the whole team, whereas here the idea was to elicit the participants' mental model of *teamwork* – or in other words, what an effective or an ineffective team would look like.

Another empirical study that dealt with measuring team model similarity was performed in the Singapore Armed Forces (Lim & Klein, 2006). Soldiers were randomly assigned to teams, and all teams received the same training program. The soldiers received training in the operations they were about to perform as a team, and also underwent extensive physical fitness training. Data collection took place at two points in time. At Time 1, 10 weeks after the teams were formed, the researchers collected survey measures of team members' and subject matter experts' task and teamwork mental models. The task mental model was defined as the team members' shared understanding of the technology and equipment with which they carry out their team tasks as well as their perceptions and understanding of team procedures, strategies, task contingencies, and environmental conditions. The teamwork mental model was defined as the team members' understanding of team members' responsibilities, norms, and interaction patterns together with the team members' understanding of each others' knowledge, skills, attitudes, strengths, and weaknesses.

The authors used their ratings to define the expert (i.e., accurate) mental models. The Time 2 data collection took place 3 weeks following Time 1 data collection, in which the team members' task mental models and teamwork mental models were measured. The researchers asked each team member to judge the relatedness (on a scale 1= unrelated, 7=highly related) of 14 statements describing team procedures, equipment, and tasks. Statements included: "Team members conducted routine maintenance of their equipment and weapons in the field"; "Team members are cross-trained to carry out other members' tasks"; "Team members have a good understanding of the characteristics of the enemy's weapons"; and "The team is highly effective." To obtain a measure of each team member's teamwork mental model, the authors asked participants to judge the relatedness (same scale) of 14 statements describing team interaction processes and the characteristics of team members (e.g., "Team members trust each other," "Team members accept decisions made by the leader," "Team members communicate openly with each other," and "Team

members are aware of other team members' abilities"). Once collected, these data were used for measuring similarity between the team members' models and accuracy, as compared to the experts' models.

TEAM MENTAL MODELS: CONCLUSION

As can be seen from the literature presented above, researchers have explored numerous techniques for team mental model and individual mental model elicitation. Many of these elicitations have led to useful predictions of team performance. Unfortunately, these techniques often require intensive researcher involvement in the data collection process and extensive further analysis after data have been collected. These elicitation processes often demand all the team members to be available and fully dedicated to the elicitation tasks. Thus, the prospects for a fully automated and unobtrusive system of mental model extraction during regular mission operations seem limited at this point, although it is certainly possible to imagine a range of future possibilities.

In contrast, the training period on the ground, with more opportunities to gather the team members together without separating one or more of them from their mission tasks, may be a more suitable period for mental model collection than an operational mission phase. In addition, if important disparities between individual or team mental models are identified at the training stage, they could ostensibly be corrected prior to actual negative impacts during operations. Moreover, mental model extraction might usefully guide the trainers in diagnosing deficiencies in a team during the training process.

Assuming that the automation and unobtrusiveness challenges could be met, repeated elicitation of team mental models during mission operations seems to have substantial potential as a method of monitoring a team and predicting future performance. Dissimilarity or lack of accuracy in mental models of the team members can be used as a warning that something in the team may not be functioning properly during the mission. It could indicate a potential for errors, decreased quality or quantity of output, or tensions between the space flight personnel due to misunderstandings.

Although mental models can expose important aspects of team members' thinking, particularly with respect to interactions with complex technologies, as currently construed in the research literature, mental models are not useful for predicting all of the outcomes of interest in the present review (e.g., outcomes such as team cohesion or morale).

Other aspects of a team member's state of mind might be obtained by analyzing the communications among the space flight personnel, between the space flight personnel and the mission control, as well as from team members' logs or any other documentation that they are required to provide. In the case of written communications, the text would be directly available for analysis, whereas for oral communications, speech to text would be used. In contrast to the mental model approach, which predicts team performance

indirectly by examining similarity among the members' models, or the accuracy of models versus an expert criterion, it may be possible to *directly* assess certain variables of interest from the contents of texts. This might allow identification of variables such as cohesion, morale, leadership, or conflicts. In the next section, we discuss a body of literature associated with natural language processing (NLP) or human language technologies (HLT) that contains various methods of extraction of variables of interest from text.

8. EXTRACTING TEAM AND INDIVIDUAL CHARACTERISTICS FROM TEXT

TEXT ANALYSIS OF TEAM MEMBER DISCOURSE

Verbal communication data gathered from members of a team can provide an indication of cognitive processing at both the individual and the team level and can be tied both to the team's and to each individual team member's abilities and knowledge (Martin & Foltz, 2004). Team communication provides a source of discourse that can be analyzed and tied to measures of team performance (Foltz, Martin, Abdelali, Rosenstein, & Oberbreckling, 2006).

In one set of studies, team communication processes were hypothesized to mediate the relationship between team member inputs and team performance (Gorman, Foltz, Kiekel, Martin, & Cooke, 2003). This research tested automatic methods that analyzed team communication in order to predict team performance. The text corpora they used consisted of team transcripts that were collected during several experiments that simulated operation of an Uninhabited Air Vehicle (UAV).

In the first study (Kiekel, Cooke, Foltz, Gorman, & Martin, 2002), three different methods were applied: Latent Semantic Analysis; PRONET (Cooke, Neville, & Rowe, 1996); and CHUMS (a method developed for this study). Latent Semantic Analysis (LSA) is a fully automatic corpus-based statistical method for extracting and inferring relations of expected contextual usage of words in discourse (Landauer, Foltz, & Laham, 1998; Martin & Foltz, 2004). This technique can measure the semantic similarity among units of text. Its "knowledge" of the language is based on a semantic model of domain knowledge acquired through "training" on a corpus of domain-relevant text. This training process uses a large corpus of text that has been meticulously tagged by human experts. The software automatically infers the rules used by the human taggers, and these rules can then be used in an automated fashion on future corpora. Because LSA can measure and compare the semantic information in verbal interactions, it can be used to characterize the quality and quantity of information expressed (Martin & Foltz, 2004).

PRONET and CHUMS represent semi-automated analytical strategies that ignore the content of communications but look at the back and forth sequencing of interactions among

different speakers. PRONET is a sequential analysis technique that relies on the network modeling tool, Pathfinder (Schvaneveldt, 1990). It is used to determine what events “typically” follow one another, after a given lag. CHUMS is a clustering tool that finds common interaction patterns and then looks for places in the discourse where pattern shifts occur. It works by clustering putative models defined by segments of the sequential data.

Using PRONET, the authors managed to identify variables that can be thought of as a measure of the team’s consistency in turn-taking behavior. Turn taking was a useful predictor of performance, primarily during early missions, when skill acquisition was still under way. For CHUMS, the researchers found that measures of team communication consistency are more predictive of performance during the learning acquisition phase of a task. The preliminary results in this study showed strong promise in using automated methods to measure team performance and cognition. Most of the methods were found predictive of performance.

A later study by these authors used LSA as the main method for analysis (Gorman et al., 2003). Within the context of a Predator unmanned aerial vehicle (UAV) synthetic task (in which skills pertinent to the corresponding real-world task can be exercised in a controlled setting), the authors developed several methods of communications content assessment based on LSA. These methods include: Communications Density (CD), which is the average task relevance of a team’s communications; Lag Coherence (LC), which measures task-relevant topic shifting over UAV missions; and Automatic Tagging (AT), which categorizes team communications. CD and LC were related to UAV team performance. The results showed that the agreement between automatic tagging and a human tagger was comparable to human-human agreement on content coding. The results proved to be promising for the assessment of teams based on LSA applied to communication content.

A subsequent study also applied LSA with the goal to measure free-form verbal interactions among team members (Martin & Foltz, 2004). In this study, the researchers used two approaches to predict the overall team performance scores: by correlating the tag frequencies with the scores and by correlating entire mission transcripts with one another. The results showed that the LSA-predicted team performance scores correlated strongly with the actual team performance measures. This suggests that LSA can be used for tagging content as well as predicting team performance based on team dialogues.

Finally, a more recent study (Foltz et al., 2006) aimed at better understanding and modeling the relationship between team communication and team performance to improve team process, develop collaboration aids, and improve the training of teams. In this study, the researchers used LSA, as well, for automating the analysis and annotation of team discourse. Two approaches to modeling team performance were described in this paper.

The first measured the semantic content of a team's dialogue as a whole to predict the team's performance. The second categorized each team member's statements using an established set of discourse tags and used them to predict team performance.

EXTRACTION OF EMOTIONS FROM TEXT

To derive the emotional state of mind of the space flight personnel, as well as the team dynamics among the team members, textual communication may also provide input for automatic text analysis to index the emotional status of an individual or status of relations among two or more individuals. As noted above with respect to LSA, the majority of techniques here use human annotated text to "train" the system or to create a model that will be able to recognize these emotions in new unannotated text.

Before creating an automated text analysis system for discovering emotions from text, Rubin, Stanton, & Liddy (2004) tried to answer the question of whether people agree in discerning the types of emotions in text, and if so, to what extent. This paper tied together a theory from social and personality psychology and NLP. The authors presented an empirically verified model of discernable emotions, Watson and Tellegen's Circumplex Theory of Affect, and suggested its usefulness in NLP as a potential model for an automation of an eight-fold categorization of emotions in written English texts. The eight categories that constitute the essence of the theory are: low negative affect (divided to subcategories such as: at rest, calm, placid, relaxed); pleasantness; high positive affect; strong engagement; high negative affect; unpleasantness; low positive affect; and disengagement. Based on the collected data, the authors concluded that the theory is useful as a guide for development of an NLP algorithm for an automated identification and an eight-fold categorization of emotion in texts.

Another study (Mishne, 2005) set out to classify various moods represented in text. Some of the moods are quite similar to the emotions in the Theory of Affect, and some are more "mood like," such as "bored." This study attempted to classify future blog posts by using existing blog text that had been classified according to the mood reported by its author during the writing. That is, given a blog post, the goal was to predict the most likely state of mind in which the post was written: whether the author was depressed, cheerful, bored, and so on. As in the vast majority of text classification tasks, a machine learning approach was applied, meaning that the task was to identify a set of features from the text to be used for the learning process. A variety of features for the classification process were used, including content and non-content features, and some features that are unique to online text such as blogs. The results showed a small, but consistent, improvement over a naive baseline. While the system success rates were relatively low, human performance on this task was not substantially better.

Strapparava & Mihalcea (2008) also utilized blogs and the moods assigned to them. The authors described the construction of a large data set annotated automatically for six basic emotions: anger, disgust, fear, joy, sadness, and surprise. The data set consisted of news headlines drawn from major newspapers. The annotators were instructed to select the appropriate emotions for each headline based on the presence of words or phrases with emotional content, as well as the overall feeling invoked by the headline. The annotators used a fine-grained scale, which allowed them to select different degrees of emotional load. For the automatic annotations, the researchers used WordNet Affect, a lexical database where nouns, verbs, adjectives, and adverbs are grouped into sets of cognitive synonyms (synsets), each expressing a distinct concept with a subset of synsets suitable to represent affective concepts. In addition to the experiments based on WordNet Affect, the authors also conducted corpus-based experiments relying on blog entries from LiveJournal.com. A variation of LSA was implemented in this study and compared in performance to UPAR7, which is a rule-based system employing a linguistic approach. The results showed that the UPAR7 system provided the best results of fine-grained evaluations, while the LSA gave the best performance in terms of coarse-grained evaluation.

The approach presented by Francisco & Gervás (2006) also used WordNet for knowledge-based expansion of words. This approach considers the representation of emotions as emotional dimensions (valence, arousal, and dominance). A corpus of example texts previously annotated by human evaluators was mined for an initial assignment of emotional features to words. This resulted in a List of Emotional Words (LEW), which then becomes a useful resource for later automated mark up. For the actual assignment of emotional features, the proposed algorithm for automated annotation employed a combination of the LEW resource, the Affective Norms for English Words (ANEW) word list (Bradley & Lang, 1999), and WordNet for knowledge-based expansion of words not occurring in either. The method for marking emotions used ideas from two of the main existing methods for marking texts with emotions: keyword spotting and lexical affinity. The algorithm for automated mark up was tested for correctness against texts from the original samples used for feature extraction and against new text samples to test its coverage. Better results were acquired for the texts used to obtain the LEW corpus than for new text.

A list of words or a dictionary was used by Frantova & Bergler (2009) as well. This paper explored automatic annotation of dream reports, which were used because they contain information that is mainly not factual, as in newspapers or scientific writing, but rather highly opinionated, sentiment-laden, and emotional. The authors compiled “emotion dictionaries” from a thesaurus using Hall/Van de Castle emotion categories proposed for dream analysis. They managed to capture the inherent ambiguity and polysemy (when a word has multiple meanings) of emotion words in word profiles, which gave a fuzzy

membership score of a word on all five emotion categories. The researchers then used the derived dictionaries to assign emotion categories to texts in so-called category profiles. The authors conclude that the system obtained good results when fuzzy category profiles were computed. Fuzziness turned out to be an inherent feature of emotions, but the observed relative ordering and strength encoded in the category profiles seems to be stable even on blog sentence sentiment annotation, a very different text type and task. In general, the comparison with the manual annotation of texts from DreamBank indicated that this multi-faceted approach was promising.

The user's emotional information was used in Guinn & Hubal (2003) to characterize his/her emotional state in interaction with virtual computer characters. This paper describes an effort to develop tagged semantic grammars that carry emotional and attitudinal information about the user's utterance. Semantic grammars are a very common form of language representation for spoken natural language processing systems. These grammars are typically domain dependent, which directly map the incoming text to underlying semantics. In addition to the semantic content of the utterance, emotional and attitudinal information was passed to the dialog manager, which utilized this information to modify its model of the user. For example, the designer of the grammar may decide that the use of the word "please" adds to the politeness of the sentence. Thus the rule would indicate that use of the rule in parsing the phrase would increase the overall sentence politeness by a small amount. Values between -1.0 and 1.0 were assigned to emotional tags. Thus a value of 1.0 for POLITENESS would be the maximum value for politeness, while -1.0 would be the most impolite phrase.

Other researchers (Zhe & Boucouvalas, 2002) have attempted to identify emotions through textual interactions such as Internet chat. These authors developed an emotion extraction engine for real-time internet text communication that could analyze input text from a chat environment and extract the emotion being communicated as well as the intensity of the emotion. Semantic analysis was used to extract emotional words. Analyzing the individual word position, the person the emotion was referred to, the time the emotion occurred, and identification of emotional words, as well as using a set of grammatical rules allowed the engine to perform satisfactorily. The engine produced better results when analyzing formal writing. Spelling mistakes and slang had a significant negative influence on the engine.

The literature reviewed above demonstrates a variety of techniques based on natural language processing for extracting team and individual characteristics from text. It appears possible to monitor certain cognitive and emotional variables through analysis of textual data generated by space flight personnel (as noted above, speech-to-text would have to be used for oral communications). Although these variables are not team-level constructs such as cohesion and morale, it may be possible to aggregate more basic emotions such as anger, disgust, fear, joy, sadness to infer team level emotions.

The following section contains a consideration of the state of the art in sensing of biological signs, or biometrics. In the context of security and assurance, a variety of sensing devices have been harnessed to assess facial expressions, galvanic skin response, infrared emanations (particularly from the face), and other markers of stress and emotion in individuals. We will be exploring whether biometrics could be used in a time-based, aggregated analysis to reveal some aspects of team functioning.

In addition, a later section focuses on proxemics, or the perception, use, and structuring of space. The research on proxemics studies how spatial use affects and reflects relationships between individuals as a member of a dyad or larger group. In that section, we describe proxemics and examine how it might be incorporated as one of the indicators for team outcomes in the context of space flight monitoring.

9. BIOMETRIC METHODS

Using and analyzing biometric data may provide another source of information to complete the full picture of how a team of astronauts functions during selection, training, or mission. In this review, we focus on emotion identification at a single moment in time, but note that the probable usage of a biometric system would be for identifying changes in emotions over time. Commercial products that employ biometric data appear to be in the early stages of development and are utilized mostly for identification and authentication purposes. The academic literature contains a range of ideas regarding the usage of biometric data, and some prototypes exist. Few commercial products exist containing a similar level of capability as these ideas and prototypes.

One type of biometric data to be considered is analysis of keystroke dynamics. For example, Andre and Funk (2005) suggest that biometrics may be used for other purposes than identification of individuals; i.e., to identify individuals' physical health status. These researchers' approach is to detect muscle tension in the users' keyboard usage, to determine the users' individual stress level. In a more recent paper, Vizer et al. (2009) reported on an initial empirical study that investigated the use of timing, keystroke, and linguistic patterns from free text to detect the presence of cognitive or physical stress. Results showed that it is possible to classify cognitive and physical stress conditions relative to non-stress conditions based on keystroke and linguistic features with accuracy rates comparable to those currently obtained using affective computing methods. The proposed approach is attractive because it requires no additional hardware, is unobtrusive, is adaptable to individual users, and is of very low cost. As mentioned above, no available commercial products were found that have a similar functionality, since most of the products that employ keystrokes are designed for security purposes (e.g., for user authentication).

A set of additional biometric measures that have been employed in several studies and prototypes are facial expressions, body movements, gestures, and speech. These indicators can be used to identify emotions either separately or in combination. Based only on speech, Ververidis and Kotropoulos (2006) presented the most frequent acoustic features used for emotional speech recognition and to assess how the emotion affects them and reviewed appropriate techniques in order to classify speech into emotional states. A combination of biometric measures was explored by Busso et al. (2004). They used a recording made by an actress who generated four types of emotions: sadness, anger, happiness, and neutral state. By the use of markers on the actor's face, detailed facial motions were captured with motion capture, in conjunction with simultaneous speech recordings. The results revealed that the system based on facial expression gave better performance than the system based on just acoustic information for the emotions considered. Results also showed the complementarity of the two modalities and that when these two modalities are fused, the performance and the robustness of the emotion recognition system improved. Similarly, Castellano et al. (2008) presented a multimodal approach for the recognition of eight acted emotional states (anger, despair, interest, pleasure, sadness, irritation, joy, and pride). This approach integrated information from facial expressions, body movement and gestures, and speech. Fusing the multimodal data resulted in a 10% increase in the recognition rates in comparison with the uni-modal systems.

Skin temperature is another biometric used in several academic articles. Khan et al. (2009), for example, employed facial thermal features in automated facial expression classification and affect recognition. A database of 324 time-sequential, visible-spectrum, and thermal facial images was developed representing different facial expressions from 23 participants in different situations. Another study that used skin-related biometric was Nakasone et al. (2005), who described a Bayesian network model that allowed determination of emotion in real time, based on electromyography and galvanic skin response signals. These two signals were chosen for their high reliability. Galvanic skin response is an indicator of skin conductance, and increases linearly with a person's level of overall emotional arousal, while electromyography measures muscle activity and has been shown to correlate with negatively valenced emotions.

Another biometric measure examined by Kruger and Vollrath (1996) used temporal analysis of speech patterns with a device named LOGOPORT, which computes the duration of four parameters for each of the two partners in conversation: (1) Undisturbed speech: when one subject is speaking and the other is listening; (2) simultaneous speech: when one subject is interrupting the other subject and both are speaking simultaneously; (3) pauses in isolation: beginning when one subject stops speaking, and ending when the subject resumes speaking, provided the second subject did not interrupt the first; (4) switching pauses: the time between speaker switching, which means the time that the second speaker

needs to take the floor. Note that these measures did not require speech-to-text conversion or any *content* analysis of the actual words spoken. Only the timing and overlap were considered. These speech patterns might be relatively easily obtained in space flight (even for multiple speakers) to gain a deeper understanding of the interactions between the team members.

As mentioned above, utilization of most biometric measures to identify emotion is currently at a research stage, and although this research might eventually be interesting and relevant for the purposes of this project, currently there are no commercial products that are capable of analyzing and utilizing biometric data for the unobtrusive detection of team states.

10. PROXEMICS

In addition to biometric indicators, we suggest exploration of a set of additional possible potential indicators – related to biometrics because they represent physical cues from a perspective of spatial relationships between people – because they may correspond to the status of relationships in dyads or larger groups. These indicators fall under the category of “proxemics” or the perception, use, and structuring of space. In proxemics research, researchers study how spatial use affects and reflects relationships between individuals as a member of a dyad or larger group, and whether it is intentional (i.e., seeking interaction) or inadvertent (i.e., in a public setting). Notable anthropologist Edward T. Hall was the first researcher who used the term proxemics. Hall developed a notation (coding) system of personal distance based on his extensive observations of humans’ use of space and evidence from animal behavior with specific reference to crowding and territoriality. Hall was particularly interested in cultural differences that appeared in people’s use of “personal” space.

Methodology in proxemics has focused mainly on interactional settings; for example, how people position themselves in a conversational setting with friends, intimates, or strangers (Harrigan, Rosenthal, & Scherer, 2008). The literature overviewed by Harrigan et al. (2008) shows that the measure coded most often in proxemics is the “distance” between the interactants. Although it might appear to be a simple task, measuring distance may be not as straightforward as it seems. A variety of different reference points have been used to represent the distance between interactants: measured from their heads, noses, knees, torsos, feet, or chair edges. This issue creates problems in the research literature because the lack of uniformity and specificity of measurement makes comparing research findings across studies more difficult. Therefore, in some studies where the independent variable was another interactant’s gender, age, culture, or personality (e.g. friendliness, dominance, inconsistency), distance was measured by the seat chosen by the participant or distance she/he approached another participant. One example of such a study is Weitz (1972) who

found that participant's chair placement reflected their attitude toward someone of a different race (2008).

While distance is an important variable in proxemics, Harrigan et al. (2008) note that it is a rather limited measure and that Hall's (1963, 1973) approach is more sophisticated and comprehensive. Hall includes the following coding variables: distance; postural identifiers (e.g., sitting, standing); orientation of frontal body plane (i.e. degree one faces another); and input from the senses of touch, vision, audition, olfaction, and temperature (e.g., perceiving heat from another's body). Hall (1963) also divided the spatial world into four social distances, each with a close and far phase, and each based on varying information available from vision, audition, olfaction, thermal reception, and kinesthesia (i.e., sensation of physical alignment of head/body). These four social distances (i.e., intimate, personal, social, and public) span zero to 30 feet, and vary according to the type of interaction and the status of and affiliation between interactants (Harrigan et al., 2008). When it comes to body movement research (kinesics), researchers' coding methods are varied, rarely well defined, and are not often organized conceptually and theoretically. Recently, researchers have been attempting to come up with a systematic coding scheme that would apply for both proxemics and kinesics, which include categories such as trunk lean, trunk orientation, arm positions, leg positions, speech illustrative gestures, self touching, object adaptors, touch, and head actions (nod, shake, tilt, dip, and toss) (Harrigan et al., 2008).

In summary, proxemics, which describes the social aspects of distance between interacting individuals, is another possible indicator to take into account. This distance represents the interactions that occur and provides information valuable to understanding human relationships (Lanz, Brunelli, Chippendale, Voit, & Stiefelhagen, 2009). Proxemics cues of importance for coding interactive behavior include: postural identification (i.e. sitting standing); distance; frontal orientation; and body positioning. Depending on research objectives, touch, eye contact, olfaction, and audition also may be coded (Harrigan et al., 2008).

In the context of this project, measuring the different proxemics variables among space flight personnel might provide indicators regarding their attitudes toward each other. Most of the research reviewed by Harrigan et al. (2008) focused on coding proxemics measures by human coders; however, more recent research demonstrated that automatic detection methods that use proxemics and kinesics to detect focus of attention (who is looking at whom), body pose, pointing, and hand-raising gestures are also becoming a viable option. To gather proxemics data, a wearable device would probably be needed (the use of the Actiwatch¹ for research purposes suggests that such devices might be acceptable to space flight personnel). While proxemics may eventually be capable of automatically providing

¹ http://www.nasa.gov/mission_pages/station/science/experiments/Actiwatch_test6.html

valuable information that helps in understanding the status and quality of dyadic relationships, there is still research to be done before proxemics could be used as predictive indicators for team performance.

11. OVERALL CONCLUSION OF THE LITERATURE REVIEW

Our review to date suggests that a substantial proportion of the industrial literature on performance monitoring may have limited applicability to space flight personnel, at least in their operational mission environments. There are probably some lessons to be learned about user acceptance of monitoring techniques, privacy concerns, and the influence of monitoring itself on motivation, but the monitoring techniques used in industrial research have been limited mainly to repetitive clerical jobs and tasks requiring a minimum of teamwork.

A promising line of research on the effectiveness of teams seeks to understand and predict team effectiveness through the elicitation of mental models held by team members. In the research, these mental models often pertain to the interaction of a team of users with a complex system. Many of the elicitation techniques described in this research are quite obtrusive and may be unsuitable except in a training environment. Automated extraction of team mental models from communicative texts is a future possibility, but one that is not extensively explored in the current literature.

In contrast, certain variables of interest have been more directly extracted from communicative texts (i.e., without assuming a mental model as an intermediate construct) using automated and semi-automated textual analysis. In one strand of research, communications among team members are analyzed to reveal either patterns of communication (as well as disruption of those patterns) or similarities and differences in the expression of various concepts. In a second strand of research, emotional states or moods have been extracted using machine learning techniques or dictionaries that encode the affective content of various words or phrases. Taken as a whole, these areas of research suggest that monitoring of individuals in teams using natural language processing or spontaneously produced communicative texts may be a viable strategy to pursue.

Finally, we presented biometric and proxemics as areas that may contain an additional set of potential indicators. Literature has shown that utilization of biometric measures to identify emotion is currently at a research stage, and although this research might be interesting and relevant for the purposes of this project, currently there are no commercial products that are capable of analyzing and utilizing biometric data. Most of the research on proxemics focuses on manually coding proxemic cues and measures by trained researchers. More recent research demonstrates that automatic detection methods that use proxemics and kinesics may become a viable option. That being said, it is important to

point out that there are limitations in the ability of biometric and proxemics measures and detection, specifically the ones that rely on visual input, especially if they will be employed in microgravity environments.

The next two sections will provide an overview of the commercial off-the-shelf products and non-commercial packages that might assist in eliciting mental models and extracting emotions based on textual communication and documents. In addition, the following material also contains summary of interviews of NASA personnel that includes their perspectives on space flight performance monitoring.

12. PRODUCTS OVERVIEW

INTRODUCTION

The literature review above focused on how individual and team performance monitoring can be achieved either with the traditional methods of industrial performance monitoring or with alternative methods that, although not designed to do so initially, may eventually enable unobtrusive acquisition of team mental models and related intrapersonal processes. These methods may provide a window into the individual or shared mental processes. Thus far, we've discussed how team mental models may be linked to measuring team outputs and what methods may be used to extract individual and team mental models. We also presented different methods of analyzing text for both mental model elicitation and for extraction of other team and individual characteristics, such as emotions.

At this stage, to provide context for an overview of the commercial and non-commercial products that enable extraction of these characteristics, figure 2 depicts a working model of several relevant precursors and outcomes. The precursors and variables are based on our interpretation of the specific competencies needed for long-duration missions mentioned in the International Space Station Human Behavior & Performance Competency Model Volume II document (Bessone et al., 2008) and on the dimensions that appeared in the expedition candidate training observation form document (NASA - Mission Operations Directorate Space flight Training Management Office, 2009). Note that this working model is likely incomplete at this stage: a number of other constructs might beneficially be included in a more mature model. Nonetheless, to evaluate available text analysis technologies, we considered it important to document our initial thinking.

Table 1 depicts the mapping between the concepts described in NASA documents mentioned above and the terms used in figure 2. The model in figure 2 is comprised of three parts: Individual Attributes; Observable Behavior; and Group States. Only the first two parts appear in the table because only the Individual Attributes and the Observable

Behavior factors are competencies, while the third part, Group States, are not competencies, but rather are emergent properties of the group or team.

Table 1: Mapping Between NASA-BHP Concepts and Working Model

General BHP Category	Competency and/or Behavioral Markers ²	Individual Attributes	Observable Behaviors
Self-care self management	“Maintains personal goals in order to feel satisfied and motivated and maximize performance”	Motivation/ Initiative	
	“Refine accuracy of self image; Identifies personal tendencies and their influence on own behavior”	Self Reflection	
Cross-cultural	“Demonstrate respect toward other cultures; Understand culture and cultural differences; Build and maintain social and working relationships; Intercultural communication and language skills; Commitment to multicultural work”	Cultural Awareness	
Teamwork and group living	“Acts cooperatively rather than competitively; Takes responsibility for own actions and mistakes; Puts common goals above individual needs; Works with teammates to ensure safety and efficiency; Respects team member’s roles, responsibilities, and task allocation”		Active Participation
	“Demonstrates effective teamwork behaviors of performance monitoring, situational awareness, back-up behavior, cooperation, coordination, information, and workload sharing”		Coordination and Monitoring
	“Volunteers for routine and unpleasant tasks”		Volunteering
Leadership	“Supports leader; Reacts promptly to situations requiring immediate response”	Loyalty	

² Based on the International Space Station Human Behavior & Performance Competency Model Volume II and on the dimensions that appeared in the expedition candidate training observation form.

On the left in figure 2, we depict some Individual Attributes that are precursors to the other constructs in the model. We define these individual attributes as inherent characteristics that each individual brings to the group. As such, these attributes comprise possible components of the astronaut selection process. Next, in the center of figure 2, Observable Behavior represents the individual and interpersonal activities that occur during the operational mission. We have chosen components that are readily manifested through either verbal or written communication. Communication serves as a window into the behaviors and, therefore, we will propose to use primarily textual communication as an input for analysis of the behaviors.

Finally, the Group States are the outcomes that team members experience as a group. As the figure suggests, Group States are the outcomes that the Observable Behaviors cause. The group states are, in all likelihood, not as readily observable as the communication behaviors but will have to be inferred from other indicators. Of course, a key goal of an automated or semi-automated system in this domain would be to predict adverse changes in one or more of the Group States in advance of their occurrence.

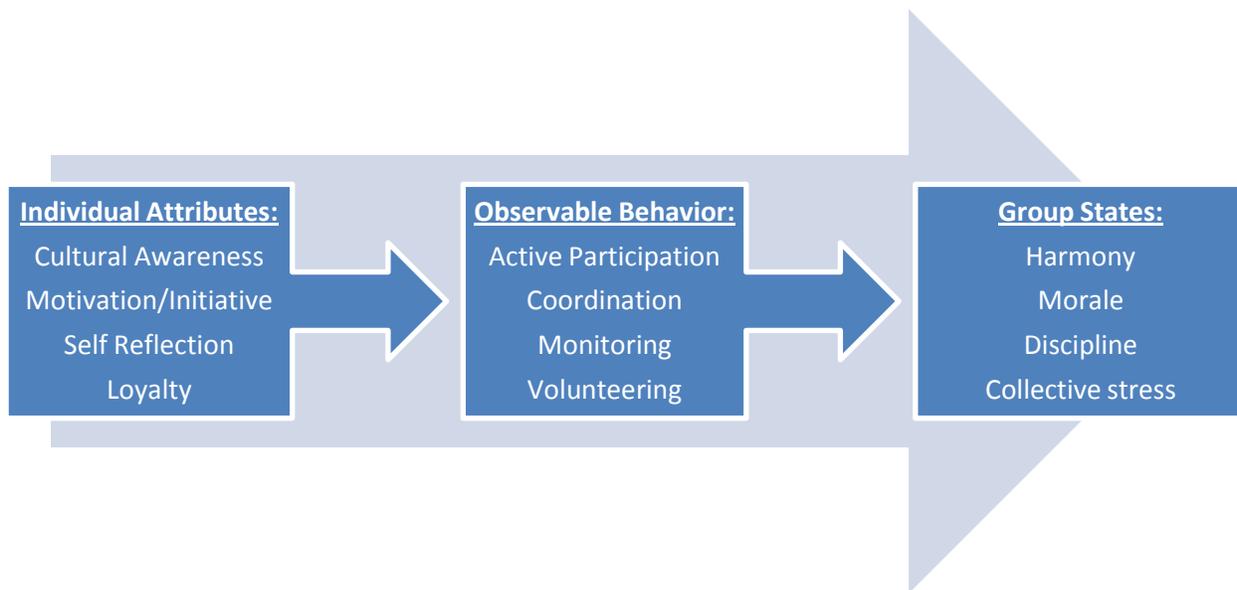


Figure 2: A working model of team functioning.

Focusing on the behavioral descriptions as they appeared in the International Space Station Human Behavior & Performance Competency Model Volume II (and also in the expedition candidate training observation form), we will briefly elaborate on the behaviors to be observed. Active team participation is expressed when a team member acts cooperatively rather than competitively, takes responsibility for her/his own actions and mistakes, puts common goals above individual needs, works with teammates to ensure safety and efficiency, and finally, respects team member's roles, responsibilities, and task allocation. Coordination and Monitoring, on the other hand, occurs when the team member demonstrates effective team work behaviors of performance monitoring, situational awareness, back-up behavior, cooperation, coordination, information sharing, and work load sharing. Finally, Volunteering is when the individual volunteers for routine and unpleasant tasks. Given this breakdown of the behavior, the main questions at this point are whether any tools or software packages can be used to identify these behaviors from communicative texts (or speech converted to text) and, once identified, whether they can be used to predict the Group States as they appear in the model above.

COLLECTING TEXTUAL COMMUNICATION

Communications among team members might provide a window into the state of mind of individuals as well as the status of interpersonal relations among team members. Text analysis methods that produce mental models may in turn provide a deeper understanding of the origin and results of team behaviors. To apply text analysis methods, there is a need to gather as much textual communication as possible. Such text-based material could be collected either during training or during a mission. Communications among the team

members, either through written messages or text obtained from speech-to-text systems, are most highly relevant, but communications between the team and the ground may also provide useful information. Likewise, individually produced texts (such as logs) might also be informative. All three sources of text may be analyzed either for mental model elicitation or for direct extraction of emotions and behaviors from text. The material below identifies the tools that may enable analysis of text produced by the space flight personnel.

A key first step in gathering spoken communication is to perform an automatic transformation from speech to text. The assumption is that recording of either audio only or both video and audio of team interactions will be available. One example of such a package is *VideoLogger* by Virage³, which uses speech recognition techniques to “watch, listen to, and read” an analog or digital video signal and create a structured video index. *Video Logger* is speaker independent, and it automatically extracts information from video data. The system is intended for semi-automated applications in storyboarding, closed captioning, and related applications. The system can also be configured to recognize faces, voices and types of sounds in the video, identify on-screen text and numbers, and convert spoken words to text. Once a video stream is indexed by *VideoLogger*, the system can be configured to automatically send an e-mail message to designated persons as an instant alert to the existence of specified information.

Dragon AudioMining by Nuance⁴ is designed to work on audio only and provides the ability to use text keywords and phrases to automatically search audio files. This software enables the indexing of 100% of the speech information within audio files. By using a speaker-independent dictation engine, it creates XML speech index data for every word spoken within an audio file. The index data includes word, time stamp, confidence levels and metadata associated with the speech information, and can be created from broadcast and telephony-quality sources. A closely related product, *Dragon Naturally Speaking*, requires speaker training (i.e., it is speaker dependent) and also requires a special dictionary if unconventional words are used.

Speech-to-text systems have several limitations and challenges. Deng & Huang (2004) found one of the challenges in developing such a system to be the ability to make it robust in noisy acoustic environments. Another challenge to be overcome is the ability to create workable recognition systems for natural, free-style speech (i.e., no pauses between words). In other words, as Deng & Huang noted, the ultimate technical challenge for speech recognition is to make it indistinguishable from the human’s speech perception. Shriberg (2005) found some features of natural free-style speech particularly problematic for speech recognition systems such as when people string together sentences without pauses, while on other occasions, people pause (as during hesitations or disfluencies) at locations

³ <http://www.virage.com/rich-media/functions/index.htm>

⁴ http://www.nuance.com/naturallyspeaking/products/sdk/sdk_audiomining.asp

other than sentence boundaries. According to Shriberg, spontaneous speech has another dimension of difficulty for automatic processing when more than one speaker is involved. An additional challenge area is to “hear” the speaker’s emotion or state of being through speech. Modeling emotion and user state is particularly important for certain dialog system applications. We will elaborate more on this type of recognition when we discuss biometric measures in the next chapter.

IBM’s *Embedded ViaVoice*⁵ is advertised to be able to deal with noise issues and continuous speech. It is available in several languages and provides both speech-recognition and speech-synthesis capabilities. The *Embedded ViaVoice* recognition engine is speaker independent because it is based on small units of speech, called phonemes. According to the developers, the maximum vocabulary supported by *Embedded ViaVoice* exceeds 150,000 words. Based on the known state of the art for other products, however, this package is unlikely to provide accurate, speaker independent recognition of such a large vocabulary.

A recently released commercial tool is Google’s automatic captioning for *YouTube*⁶ videos. Their system is also designed to deal with free-style speech in the presence of environmental noise, because these are typical characteristics of YouTube videos. No public evaluation results are available for this system; therefore, it is difficult to assess the quality of this recent speech recognition system. Anecdotally, the system provides a minimally useful first approximation of the spoken text that must be edited for accuracy by a human user.

One may bypass the speech-to-text data collection phase by focusing on textual messages created on a keyboard or related input device. E-mail or computer monitoring software may be used for the purpose of gathering and aggregating all the text that has been typed on a computer. Commercial e-mail monitoring software is intended mostly for surveillance of actions performed by employees on their work computers. Both online and offline activities can be recorded and then reproduced by the employer for viewing and analysis. After reviewing several of these software packages, we found that most e-mail monitoring software functionality is geared toward detailed monitoring of each employee separately. The software does not provide any tools for in-depth analysis to get the meaning behind the text, but instead it suggests mostly simple presentation of keyword frequencies or alerts on keywords defined by system administrators. In other words, the reports generated by the systems allow only a very shallow form of text analysis. Therefore, this type of software may be useful mostly for text gathering, for further analysis by other tools or software packages.

⁵ http://www-01.ibm.com/software/pervasive/embedded_viavoice/about/?S_CMP=wspace

⁶ http://www.wiredprnews.com/2010/03/05/youtube-expands-automatic-caption-feature_201003059289.html

One example of such software is *Spector 360*⁷. This package allows very detailed monitoring that enables the employer to see what an employee does each and every second of the work day. It also allows generation of reports and charts across employees to help identify those employees who are most likely engaging in activities that are harmful to the company. A system administrator can define keywords that will be extracted from the text typed by the employee and used to generate charts or alerts that will appear in the report. The main feature relevant to this project's case is the ability to record keystrokes. It includes a keystroke logger that saves keystrokes by application, date and time, and also who typed it, based on user login information. The package also records "hidden" characters and keystroke combinations, such as the Shift and Ctrl key.

Another package that performs keystroke logging is *Keystroke Spy*⁸. This product also a monitoring solution that can log every keystroke and that captures screenshots of everything they do. It also has an option of delivering alerts when any of a list of keywords is typed. It is capable of logging everything that is typed on the computer or alternatively logging keystrokes typed in specific applications and windows.

Other monitoring tools are capable of alerting not only for specific predefined keywords, but also for predefined patterns. An example of such software is *Mimecast*⁹, which includes regular expression testing¹⁰. This feature's goal is to detect variable data within the text, for example data such as Social Security and credit card numbers or any other data that fits a certain pattern. Although still a very shallow form of text analysis, it could be useful for certain basic detection purposes. For example, there are libraries of words that express extreme affective content (e.g., hate, despise, adore, ecstatic) that could be deployed with this type of software to provide a rudimentary system for flagging messages that express strong emotions.

Once all the communication has been collected, whether it originated in textual form or was transformed to textual form, sophisticated textual analysis may be applied on the text to detect entities, relations, patterns, and other higher order structures. There are no commercial packages that were explicitly designed for the purpose of eliciting mental models, but using text analysis for this purpose or for the purpose of recognizing the existence of a specific behavior might still be feasible, either with a custom-developed system or with off-the-shelf systems available in the near future.

⁷ <http://www.spector360.com/>

⁸ <http://www.spytech-web.com>

⁹ <http://www.mimecast.com/email-monitoring/>

¹⁰ In computer programming, a "regular expression" is a method for describing a variety of letters, words, or phrases that fit a user-specified pattern. For example, the regular expression "colou?r" could be used to match the alternative spellings color or colour.

TEXT ANALYSIS PACKAGES

To examine whether it would be possible to gain insights into individual and team processes in an unobtrusive manner, we reviewed the possibility of applying text analysis to the communications generated by team members in conversation with each other as well as with personnel on the ground. The literature review exposed a family of systematic content analysis methods intended for analyzing written statements such as formal speeches and transcripts of interviews. This is the most unobtrusive method among all the mental model elicitation techniques, and therefore, text analysis software that allows performing content analysis was of particular interest. With proper modeling and “guidance,”¹¹ this type of software may also be capable of extracting emotions and behaviors based on the text. We looked at several commercial off-the-shelf products that offer textual analysis, usually within an organizational or research context. Below, we present a table summarizing a set of software packages and tools that perform various forms of text analysis, along with their advantages and disadvantages in the context of this project.

Table 2: Advantages and Disadvantages of Text Analysis Software

Name of the Software/Tool	Advantages	Disadvantages
<i>BusinessObjects Text Analysis</i>	<ul style="list-style-type: none"> - Extracts information from unstructured text sources such as e-mails, web pages, and documents. - Provides alerts to new or changing information as it develops and allows navigation between relationships, concepts, and timelines. 	Intended for general business purposes; requires a human operator to interpret the results and perform further analysis.

¹¹ In the research area of human language technologies, as well as in related research areas, it is common practice to divide a data set, such as a corpus of text, into a training set and an evaluation set. The training set is used to “train” the software to recognize or detect certain patterns or relationships. Following completion of the training, the software then attempts to recognize the patterns or relationships in the evaluation set. The performance of the software is then gauged by comparison to a known (or derived) statistical benchmark or the results produced by human experts.

Name of the Software/Tool	Advantages	Disadvantages
<i>PolyAnalyst</i>	<ul style="list-style-type: none"> - Allows knowledge discovery in large volumes of textual and structured data. - Enables intelligent analysis of data and text by producing easy-to-understand actionable results. - Allows incorporating dictionaries such as Wordnet. 	Operates through interactive drill down and visualizations, all of which require human operators.
<i>TextAnalyst</i>	<ul style="list-style-type: none"> - Capable of distilling the semantic network of a text completely autonomously, without prior development of a subject-specific dictionary by a human expert. The user does not have to provide any background knowledge of the subject – the system acquires this knowledge automatically. 	Operation both on the input side and on the output side is required by a human user.
<i>PASW Text Analytics for Surveys 3.0</i>	<ul style="list-style-type: none"> - Allows analyzing open-ended questions on a survey by employing text analysis and visualization methods. - Quantifying text responses for analysis and automates the process while enabling to intervene manually to refine the results. 	Visualization results can be interpreted only by human intervention.
<i>Attensity</i>	<ul style="list-style-type: none"> - Employs semantic approaches to extract and recall information hidden in free-form text, turning it into insights that can be used by all types of business users. - Fusing unstructured and structured data provides an overall picture of the data. - The technology allows users to extract and analyze facts like who, what, where, when, and why, and then allows users to drill down to understand people, places, and events and how they are related. 	Intended for business intelligence purposes, the output needs to be manipulated by human users.

Name of the Software/Tool	Advantages	Disadvantages
<i>Diction 5.0</i>	- This software has a specific purpose of identifying affective tone in a verbal message by performing text analysis.	Only some of the affective tones analyzed by this software are suitable for the purposes of this project.
<i>LIWC 2007</i>	- Capable of detecting emotions and other dimensions in unstructured data.	Word usage in the text needs to be rather explicit for the software to detect the contrast between positive emotions and negative emotions.
<i>KNOT</i>	- This software is built around the Pathfinder network generation algorithm, which is a technique to elicit individual and team mental models.	The requirement of the input to be processed and presented in a form of comparison data.

*BusinessObjects Text Analysis software*¹² by SAP extracts business information from unstructured text sources such as e-mails, Web-based, and customer documents. The vendor suggests using this software to analyze customers, root causes, links, shareholder value, counterterrorism, or employee satisfaction. Main features of this software include entity extraction and analysis, taxonomy-based document categorization, and automatic document summarization. The package can analyze data over time and report on dynamic changes to variables it derives from the data. Once information is collected, the extraction and analysis tools in the software allow navigation of relationships, concepts, and timelines. This software structures language into its most basic parts through automatic language and character encoding identification, document analysis, word segmentation (tokenization), stemming, normalization, decompounding, part-of-speech tagging, and noun phrase extraction.

A similar software package is *PolyAnalyst*¹³ by a firm called “Megaputer” (please see the hands-on product review at the close of this chapter). *PolyAnalyst* is a tool for knowledge discovery in large volumes of textual and structured data. This system was designed with the goal of enabling firms to answer business questions by scanning unstructured historical data and predicting outcomes of future situations through interactive drill down and visualizations. The interface offers analysis tasks including categorization, clustering, prediction, pattern learning, trends analysis, anomaly detection, link analysis, entity extraction, natural language search, and graphical multidimensional reporting.

¹² <http://www.sap.com/solutions/sapbusinessobjects/large/information-management/data-integration/textanalysis/index.epx>

¹³ <http://www.megaputer.com/polyanalyst.php>

Another product from Megaputer is *TextAnalyst*¹⁴, which helps users deal with large amounts of text. *TextAnalyst* is intended to summarize, navigate, and cluster documents in a textual database. It can also provide the ability to perform semantic information retrieval or focused text exploration around a certain subject. Specific functionality includes:

- Distilling the meaning of a text – formation and export of a Semantic Network of the text. A Semantic Network is a set of the most important concepts from the text and the relations between these concepts weighted by their relative importance. This network concisely represents the meaning of a text and serves as a basis for all further analysis.
- Summarization of texts – performed by utilization of linguistic and neural network investigation methods. Allows controlling the size of the summary.
- Subject-focused text exploration – user-specified dictionaries of excluded and included words allow the investigation to focus on a chosen subject.
- Navigation through a textual database – the knowledge base can be navigated with hyperlinks from concepts in the Semantic Network to sentences in the documents that contain the considered combination of concepts.
- Explication of the text theme structure – a tree-like topic structure representing the semantics of the investigated texts is automatically developed. The more important subjects are placed closer to the root of a tree.
- Clustering of texts – breaking links representing weak relations in the original Semantic Network enables clustering of the textual database.

In a more research-focused domain, *PASW Text Analytics for Surveys 3.0*¹⁵ by SPSS was created for the purpose of analyzing open-ended questions on a survey by employing text analysis and visualization methods. Although intended for surveys, it is possible to imagine applying its textual analysis functionality for the purposes of mental model or behavior extraction. The package allows quantifying text responses for analysis and automates the process while at the same time enabling to intervene manually in order to refine the results. The package's main capabilities include identifying major themes, distinguishing between positive and negative phrases, extracting key concepts and opinions, summarizing findings, creating and applying categories, and exporting results for analysis and graphing.

A text analysis package designed for business purposes is *Attensity*¹⁶. This software has several modules, and the most relevant for this project's purposes are the *Semantic Engines*¹⁷ module and the *Text Analytics*¹⁸ module. The first module employs semantic approaches to extract and recall information in free-form text. The interface allows users to explore the relationships between topics, without having to manually read the whole

¹⁴ <http://www.megaputer.com/textanalyst.php>

¹⁵ www.spss.com/media/collateral/data-collection/STAS3SPC-0509.pdf

¹⁶ <http://www.attensity.com>

¹⁷ <http://www.attensity.com/en/Technology/Semantic-Engines.html>

¹⁸ <http://www.attensity.com/en/Technology/Text-Analytics.html>

corpus. The software provides keyword search, classification, clustering, categorization, machine learning, case-based reasoning, name entity recognition, language identification, event and relationship extraction, and artificial intelligence. On the linguistics side, it provides exhaustive extraction, advanced pattern recognition, and semantic web.

The *Text Analytics* module automatically extracts data from free-form text. The technology allows users to extract and analyze entities, relations, and events over time. Premade “schemas” are available; these provide aggregated data views that support the schema formats for most of the business intelligence applications in the market.

Most of the tools presented above don’t have a specific analytic goal, but rather are intended to be applicable for a variety of business and research activities. They all assume that a human operator is involved in the process – either on the input side, the creation of the model stage, or on the output side. These packages are incapable of reaching a conclusion or recommending an action; a human user needs to use the output of such software in order to make an informed decision.

In contrast, the following software may be one step closer to one of this project’s goals, which is to extract emotion from text automatically. This software, called *Diction 5.0*¹⁹, is a software package that performs text analysis for the purpose of determining the tone in a verbal message. *Diction 5.0* uses dictionaries (word-lists) to search through text for the following qualities:

- Certainty - Language indicating resoluteness, inflexibility, and completeness and a tendency to speak authoritatively.
- Activity - Language featuring movement, change, the implementation of ideas, and the avoidance of inertia.
- Optimism - Language endorsing some person, group, concept, or event, or highlighting their positive entailments.
- Realism - Language describing tangible, immediate, recognizable matters that affect people's everyday lives.
- Commonality - Language highlighting the agreed-upon values of a group and rejecting idiosyncratic modes of engagement.

Another software that was designed to detect emotion, as well as other dimensions, is *LIWC*²⁰ (Linguistic Inquiry and Word Count), which analyzes written or transcribed verbal text files by looking for dictionary terms matched to words in the text. It is done on a word-by-word basis by calculating the percentage of words in the text that match a particular dimension in the dictionary (Sexton & Helmreich, 2003). LIWC includes several dimensions such as linguistic (pronouns, first person, articles, prepositions, etc.), psychological process

¹⁹ <http://www.dictionsoftware.com/>

²⁰ <http://www.liwc.net/>

dimensions (positive emotions, negative emotions, cognitive processes, and so on), and more. The software employs dictionaries comprised of words that represent each dimension (for example: the positive emotion dimension is represented by keywords such as happy, pretty, good, and other positive terms). It calculates the percentage of words in a section that fall into each dimension.

An additional software program dedicated to a specific text analysis objective is *Knowledge Network Organizing Tool (KNOT)*²¹. One of the methods to extract and measure team and individual mental models is Pathfinder (PF, also mentioned in the literature review), which is intended to produce psychological scaling of the underlying structure between concepts. The PF algorithm transforms raw, paired comparison data into a network structure in which the concepts are represented as nodes, and the relatedness of concepts is represented as links between the nodes. *KNOT* is built around the Pathfinder network generation algorithm. Pathfinder algorithms take estimates of the proximities between pairs of items as an input and define a network representation of the items. The network (a PFNET) consists of the items as nodes and a set of links (which may be either directed or undirected for symmetrical or non-symmetrical proximity estimates) connecting pairs of the nodes. The set of links is determined by patterns of proximities in the data and parameters of Pathfinder algorithms. The system is oriented around producing pictures of the solutions, but representations of networks and other information are also available in the form of structured text files that can be used with other software. The disadvantage of this system is in the requirement of the input to be processed and presented in a form of comparison data. This means that utilizing raw text from communication requires another stage of processing, performed by a human expert.

In summary, the software reviewed above exhibits promising capabilities to transform unstructured text into useful visualizations and other analytic output. These packages provide the opportunity for a human analyst to obtain a sophisticated understanding of a large corpus of text. At this writing, there is no purpose-built software that will process a corpus of text obtained from one or more sources and automatically extract from that text a mental model or other high level construct. As can be seen from the disadvantages presented in Table 2, text analysis may be the closest automatic method to mental model elicitation that is also commercially available. At the same time, this software is still not the perfect tool for unobtrusive acquisition of mental models, emotions or behaviors from text. Text analysis requires a human operator to look at the results, further analyze, and interpret them.

²¹ <http://pathfindernets.com/KNOT.html>

13. A CASE STUDY OF *POLYANALYST* SOFTWARE

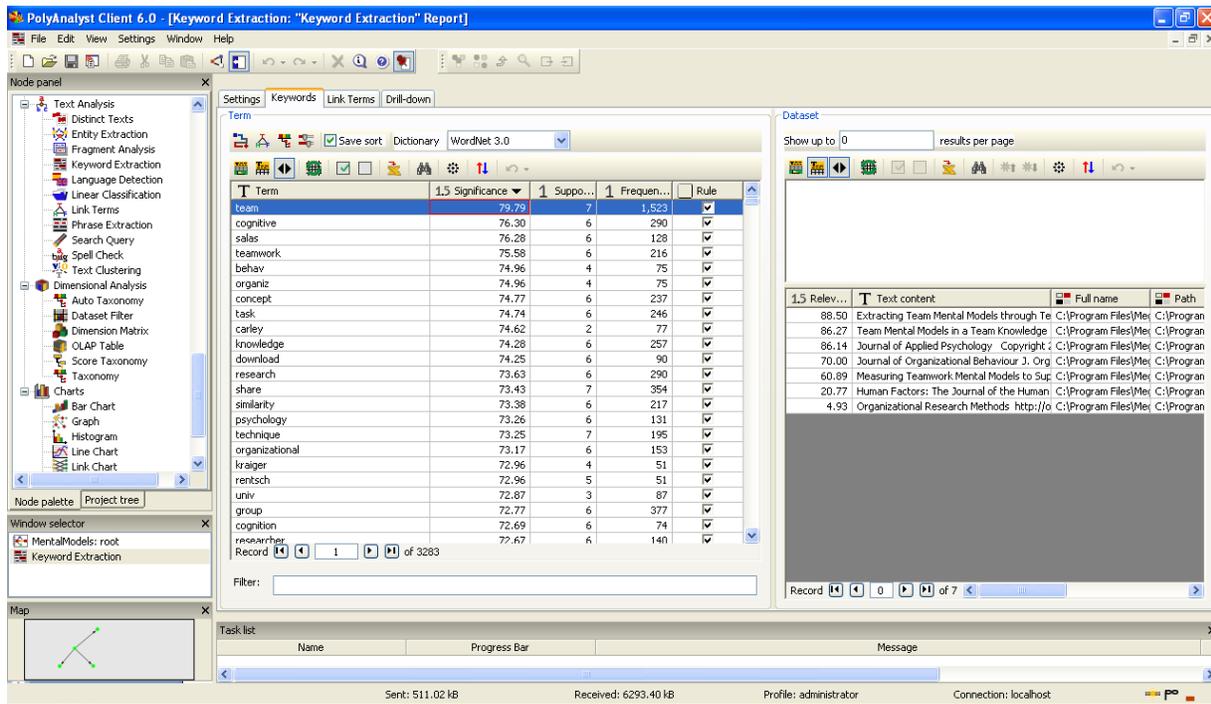
To demonstrate how the family of text analysis software packages operates, we performed a very brief exploration of the functionality of *PolyAnalyst*, one of the text analysis packages. This software enables creating a flow of nodes that feed one into another. In other words, the output of one block may be the input of another block, the types of nodes being: data sources, column, row, and table operations, as well as data analysis, text analysis, dimensional analysis, and charts (visualization) functions. The graphical user interface of this software allows using drag-and-drop to choose from the list of available nodes and make the connections between the nodes. Following are a few examples of potential utilization of several *PolyAnalyst* functions.

Given test data in a form of nine PDF files (containing text from academic papers about mental models) the following functions were used to analyze the text in those files: phrase extraction²², keyword extraction²³, and auto taxonomy²⁴. The primary output of the Keyword Extraction node is a report displaying keywords and information about keywords. For each word in the report, the significance, support, and frequency are listed. The significance is a calculated measure that describes how unique and distinct that keyword is to the current text being analyzed. The support is the number of records that contain the keyword. The frequency is the number of times the keyword appears in all the files. The following screenshot demonstrates how the Keyword Extraction report appears.

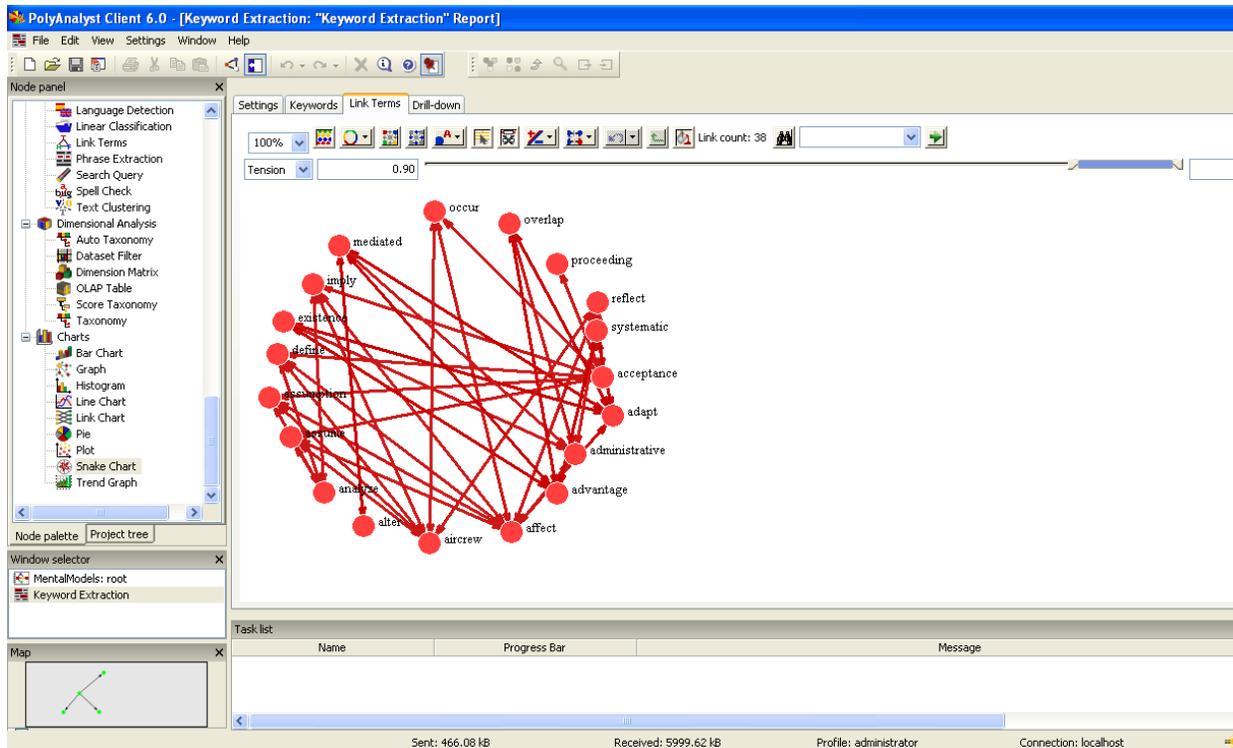
²² The phrase extraction process derives phrases (a group of alphabetical words that occur next to each other within natural language) statistically by examining the co-occurrences of consecutive words within the text. If two words occur next to each other repeatedly in several sentences across several documents, it can be statistically assumed that these words constitute a phrase.

²³ Keyword extraction derives keywords that are unique and distinct in the current text base being analyzed.

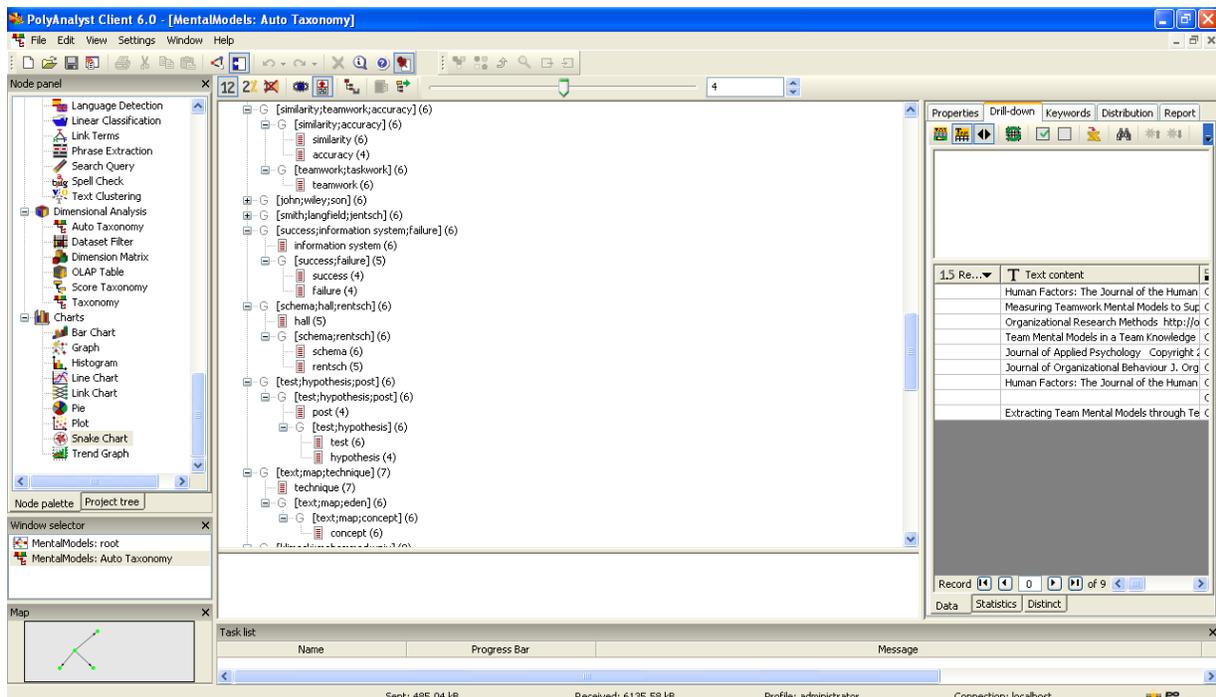
²⁴ A taxonomy usually has a hierarchical structure similar to a tree. It starts with a root category and underneath the root category there are one or more subcategories (the branches) and possibly more subcategories underneath those. The categories at the very bottom of the tree are often referred to as leaf categories. Auto-taxonomy is derived automatically from the text, based on the co-occurrences of the words.



From the Keyword Extraction report, users can view drill-down results by clicking on a keyword, or drilling down the Link Terms graph and revealing which text contains the selected keyword. The Link Term graph is a visualization that is also part of the Keyword Extraction report. As can be seen in the picture below, this Link Terms option displays a graph of correlated keywords and phrases. By increasing the minimum threshold, the users can filter out relations that have very little support (a low number of records where the two words appear). By decreasing the maximum threshold, the user can filter out some of the very obvious relations (like words that constitute phrases and are always mentioned together).



A feature called Auto Taxonomy presents a taxonomy generated automatically from the text. The tree of categories is displayed on the left, with a count of records matching each category. Similar to the previous functions, clicking on every category in the taxonomy enables drill down to the specific text or records that contributed to the creation of that branch.



As the demonstration suggests, the functionality of this software and other similar packages is usually preset to a limited group of functions. Some of these functions may support the goals of this project to elicit behavior and relations between individuals from text, but this would require turning collected texts into a more structured form.

The output produced requires a human expert to interpret it in order to make an informed decision. Although there are several drawbacks, *PolyAnalyst* does have a few visible advantages that might make this software promising for the purposes of this project. First, among the various methods that can be used for mental model elicitation, the closest method that can be utilized within a software package is content analysis. Instead of using coding rules for content analysis, general patterns can be extracted from the text and keywords are used to represent the text. This software is capable of supporting these functionalities and also enables exploring the links between concepts and keywords within the text. *PolyAnalyst* allows incorporating various dictionaries (WordNet for example) and machine learning functionalities such as classification, which eventually may lead to the ability to extract emotions and behaviors from text. The workbench style of this software implies that this software enables constructing complex processes that may be operated semi-automatically.

Given the brief exploration above, more information (for example how to transform the text to a more structured form) and further exploration are needed regarding the full potential of this software and its application to the goals of this project. In the past, the potential of this particular software has already been explored by other companies, such as Southwest Airlines, who performed a proof-of-concept demonstration of data and text mining in order to facilitate and promote the use of automated data and text mining for improving overall flight safety performance (Ananyan, Kasprzycki, & Kollepara, 2004). This project proposed new techniques and methodologies to conduct analysis of flight safety data to reveal associations and trends that may otherwise be difficult and time consuming to identify. Although the data used by Southwest Airlines is slightly different and more structured than the data available in this project, parallels may be drawn between the two cases and provide reassurance for further exploring the capabilities of *PolyAnalyst* for this project. Since other institutions in the aviation community have considered this software, it might have promising potential for this project as well.

14. A CASE STUDY OF LIWC SOFTWARE

We chose LIWC software program for this demonstration because it was previously deployed in a study conducted in a setting similar to the one we are investigating. This study, by Sexton and Helmreich (2003), explored the use of language in the cockpit and examined its relationships with workload and performance. The authors chose to study communication within the air crew, because previous research had shown that crew

performance was more closely associated with the quality of crew communication than with the technical proficiency of individual pilots or increased psychological arousal as a result of higher workload. Sexton and Helmreich used cockpit communication data that was originally collected for an investigation of the effects of captain personality on crew performance. These data were derived from transcribing four flight segments that involved a three person crew (captains, first officers, and second officers) flying a simulated Boeing 727 during a five-segment flight over 2 days. As part of the original data collection, an expert pilot observer was present in the simulator and recorded data regarding individual performance, individual errors, and individual communication skill.

In their study, Sexton and Helmreich found that word count (overall number of words spoken), first person plural (“we”), and number of questions asked in the first flight were positively related to performance and communication as well as negatively related to rates of error. A similar pattern was found for use of the present tense and discrepancy words (“would, should, could”). They also found that captains consistently used more words, used more first person plural, and asked fewer questions than the other crewmembers. Captains also used more present tense than first officers and second officers. The authors presumed that present tense usage is a marker of verbalization such that pilots, who verbalize their actions more, use more present tense, and that linguistic dimension is related to flight outcomes (individual performance, individual errors, and individual communication skill). They also inferred that pilot’s use of discrepancies could be an indicator of linguistic politeness in the cockpit and that a pattern of increasing use of the first person plural might indicate an increasing sense of familiarity among the crewmembers or an increase in their team perspective.

To demonstrate additional capabilities of LIWC, we analyzed two Wikipedia entry discussions. A Wikipedia discussion is a dedicated page assigned to every Wikipedia entry, which displays the comments of Wikipedia contributors regarding proposed changes in the contents of that particular entry. The discussion is comprised of comments posted by the contributors in an attempt to settle an issue or disagreement that the contributors have concerning different parts of the entry. The first entry that we chose for this demonstration was a description of a certain event in history, and we analyzed the contributor discussion on how it should be labeled. There was a disagreement among the contributors regarding the labeling of the event because it was politically charged, and therefore a discussion was started on the discussion page. The discussion led to a rather heated exchange of comments fueled by the differences in political perspectives of the contributors. The other Wikipedia entry that we chose to analyze describes a city in New York State, is not charged politically, and has a much “friendlier” discussion around it.

We fed the data from the Wikipedia discussions into LIWC. Each comment was considered as a separate section, and for each section LIWC calculated the percentages (scores) that

fell into each dimension. We also edited the dictionary in order to match it to the specific topic and the nature of textual communication. Specifically, “thank you” and “please” were removed from the positive emotion dimension dictionary because these words are usually used as a matter of formal politeness in this type of communication and do not actually reflect positive feelings. In addition, the word “attack” was also removed because it was part of the subject matter being discussed in the comments of the first entry and therefore was used not necessarily to express negative emotions among the contributors. Each comment in the discussion was analyzed separately by LIWC, and thus LIWC calculated a different score for each dimension and each comment. This score reflects the percentage of words in the text that matched the keywords of a specific dimension. For the sake of this demonstration, we looked specifically only at the negemo (negative emotion) and the posemo (positive emotions) dimensions, because they are the most relevant for our study, and we used the scores generated by LIWC to create a box plot.

Box plot²⁵ is a graphical method for representation of a set of data points. In our case, the value (y axis) is the score that LIWC provided based on the percentages that fell into each dimension. The median of the values is identified by a line inside the box. The body of the box plot consists of a “box,” which stretches from the first quartile (the 25th percentile) to the third quartile (the 75th percentile). Two lines (whiskers) extend from the upper edge (top) and the lower edge (bottom) of the box. The upper whisker goes from the top of the box to the largest non-outlier in the data set, and the lower whisker goes from the bottom of the box to the smallest non-outlier. Outliers are marked as small circles on the plot and signify data points that differ greatly from the overall pattern of data.

Our set of data points was based on LIWC output; each data point was a comment in the discussion of the first Wikipedia entry or in the second Wikipedia entry. Figure 3 represents the data points derived from the first entry discussion (on the labeling of a historical event) and Figure 4 represents the data points derived from the second entry discussion (on a city in New York State). As can be seen from the figures below, LIWC was able to identify the differences between the emotions expressed in each of the entries. For both figures, the “0” category represents the negative emotion dimension and the “1” category represents the positive emotion dimension. It can be seen that for the first entry, the one with a more hostile discussion, the percentage of words that express negative emotions (“0”) is generally higher than the percentage of words that express positive emotions (“1”). For the second entry, the one that had a “friendlier” discussion, it can be seen from Figure 4 that the percentage of words that express positive emotion (“1”) is generally higher than the percentage of words that represent the negative emotions (“0”).

²⁵ <http://stattrek.com/AP-Statistics-1/Boxplot.aspx>

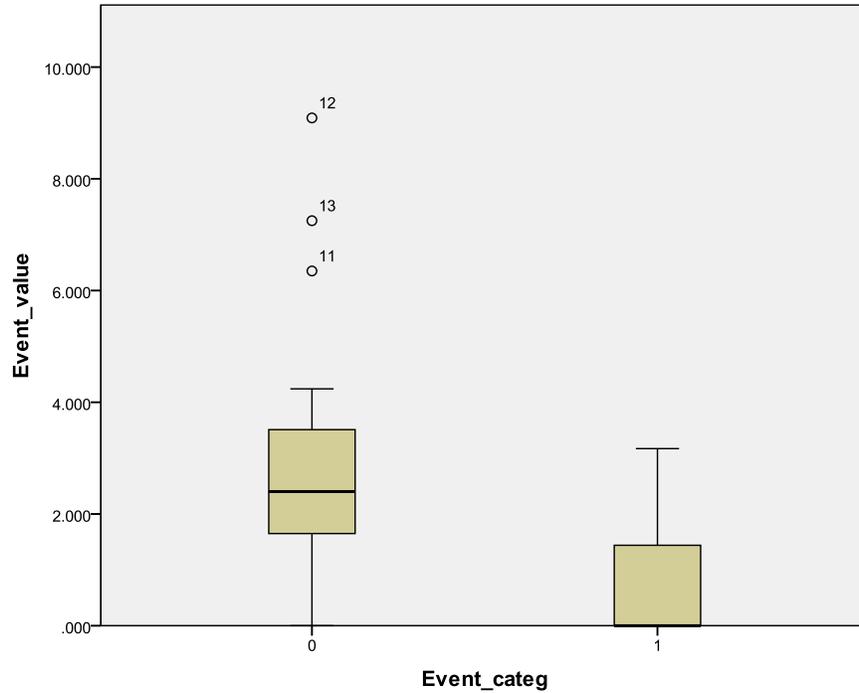


Figure 3: Box plot of LIWC output for the historical event entry discussion.

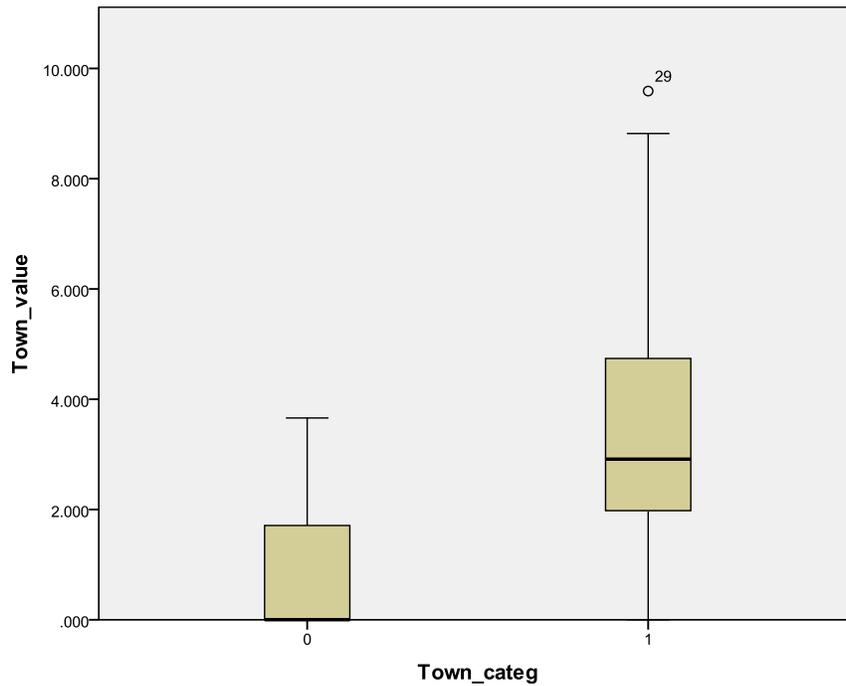


Figure 4: Box plot of LIWC output for the discussion on the city entry.

In summary, LIWC can be a useful tool for easily discovering linguistic and emotional dimensions from transcripts and can potentially add more insight into what takes place in the minds of the team members. The output of the software can be made more accurate if the dictionary is adjusted to the nature of the text being analyzed. It is important to note

that based on different attempts and experimentation with various types of text, we have come to the conclusion that in order for the software to be able to discern between the dimensions, especially those associated with emotions, the text needs to contain significant linguistic contrast and explicit wording that is typical to these dimensions.

15. SUMMARY OF INTERVIEWS WITH KEY PERSONNEL

We interviewed four people who are currently involved in observing or monitoring space flight personnel or are in charge of the technology that allows communication and monitoring. The objective of these interviews was to elicit their perspectives on monitoring team performance during long-duration missions and the feasibility of a potential automatic monitoring system. Our interviewees were: an On-console Communication and Tracking Officer, an Aerospace Psychologist, an Operational Psychologist, and a Flight Controller. Following the constraints of our institutional review board research approval, we withhold the identities of these individuals. For the same reason, we have refrained from presenting verbatim quotes. The following is a synthesis of some of the inputs we got from our respondents.

Currently, 24-hour video transmission is not provided from the International Space Station. Availability of video depends on permission given by the astronauts. For public affairs events, high-definition video and audio are transmitted to the ground, but for all other purposes, the video is in standard definition, and the level of audio quality is variable. Noise on board might be an issue with respect to picking up dialog among the astronauts, and monitoring technology will need to compensate for that. Besides the traditional means of communication with the ground, the astronauts also have email and twitter access. A suggestion of using gyroscopes in order to identify human movement on the station would likely be superseded by more direct means of assessing personnel activity, such as the “Actiwatch” mentioned previously in this review.

While there are no joint international operations, personnel often reside in their own modules and have lesser levels of interaction with their international colleagues. This might be due to the fact that sometimes the astronauts went on a mission after they had seen the other members of the crew (especially the international ones) only a couple of times. Retreating to their modules may also indicate social frustrations among the astronauts. Observing the dynamics between the astronauts from different cultures might be of value because there are some culture-specific traits that may lead to conflicts (in spite of extensive training). Some ground personnel have noticed a “curve” of progress in social relations, where the astronauts’ liking increases as time progresses, then levels off, and then improves again near the end of their mission period.

When asked about indicators that would suggest problems in team functioning, our interviewees mentioned fatigue, which may be expressed by limited conversation among the team members. On the other hand, a lively discussion that turns into a heated argument may also be an indicator of problematic team functioning. Usually, when things are going poorly, the crew will bring it up in one of the private sessions that the astronauts have with their psychologists and complain about irritation, not getting along with someone, personality clashes, or team dysfunctions.

In terms of assessing team performance, the interviewees suggested examining the type of interactions, inquiries, questions, asking for assistance, how many declarative statements are made, how many coordinating statements are made, how many interactions occur per minute, how quickly a team gets the task completed, and instances where the astronauts are not following procedures. Some crews on shuttle flights have higher error rates than others. A number of tasks are more critical than others, and the tasks that require more focus and might be dangerous are: launch, docking, ISS-Soyuz relocation, robotic arm operations, extravehicular activities coordination with crew inside and outside, and landing.

Sometimes mission control personnel have observed the dynamics between the astronauts in body language, posture, and distance between the crew members. Ground personnel try to detect deviations from the crew's usual behavior. They pay attention to open air to ground channels, such as when sarcastic remarks about the space or ground crew are couched in humor. Some tensions may be evident from the videos transmitted to the ground, for example an incident of microphone grabbing between the crew members during a public affairs event. One of the interviewees noted that the best way to predict problematic team performance so far has been to gather feedback from the crew itself through verbal self report. This respondent incorporated a self report system with his ground crew and inferred from this process that as long as personnel feel protected from the upper management, they do not hesitate to report their mistakes and suggest what can be improved.

We asked interviewees about reactions to monitoring as well. We learned that one of the reasons why video is at the astronauts' discretion is because they don't want to feel critiqued. Also, when the cameras are on, the astronauts will try to present themselves as likeable and task oriented, because they want to be assigned to more missions. Historically, in earlier phases of space activity, performance led to rewards or punishment, and this has resulted in a reluctance to be monitored. Astronauts want to come across as very confident and not to show that they might be having difficulties. To improve astronauts' reactions to monitoring, astronauts must buy into the importance of the monitoring system and be shown that its advantages are significant, such as reducing errors, or increasing time efficiency or safety.

Overall, the interviews provided us with insights regarding the current state of monitoring the astronauts and their potential reactions to automatic monitoring, as well as what we should take into consideration or pay more attention to when designing a monitoring system. We learned that negative reactions to monitoring may exist, and that there is a need to convince the astronauts of benefits from monitoring as well as to ensure that they will be protected from adverse uses of monitoring data. In addition, we confirmed that other aspects such as body language, proximity between the astronauts especially from different cultures, and deviations from their normal behavior could be considered as indicators.

16. OVERALL CONCLUSIONS

The literature review portion of this document showed that the research on industrial performance monitoring has limited value to space flight operational mission settings. The review suggested that a more relevant line of research exists focusing on the effectiveness of teams and how team effectiveness may be predicted through the elicitation of individual and team mental models. Note that the “mental models” referred to in this literature typically center on a shared operational understanding of a problem space, such as the cockpit controls and navigational indicators on a flight deck. In principle, however, it is not difficult to imagine that such mental models exist, reflecting the status of interpersonal relations on a team, collective beliefs about leadership, success in coordination, and other aspects of team behavior and cognition. Although many of the elicitation techniques described in the literature review are quite obtrusive, and may be unsuitable except in a training environment, the second part of this document provided an overview of the available off-the-shelf products that might reflect future possibilities for extraction of mental models and elicitation of emotions based on the analysis of communicative texts. Another possibility explored in this document is the option of incorporating biometric measures in order to expose various individual states (such as stress) that may be indicators or predictors of certain elements of team functioning.

The search for text analysis software or tools revealed that currently there are no available commercial off-the-shelf tools that enable extraction of mental models automatically and unobtrusively, relying only on collected communication text. Commercial text analysis software is, on the one hand, too general and, on the other, not flexible enough to be operated without human intervention. Therefore, usage of this software to derive how a team is functioning and what its mental models are may be relevant for the selection or training stages, when human operators are available. Alternatively, if output from the software described above or from a modified version can be sent to the ground periodically and analyzed by experts on the ground, then these software packages might be employed during missions as well. Clearly, since the packages and tools reviewed in this document

were designed mostly for business purposes, utilizing them as-is will not be optimal, and adaptations to the space flight context will be required. Nevertheless, the core capabilities of these packages may be useful as a starting point.

In addition, emotion detection software applications may be useful tools to easily discover linguistic and emotional dimensions from transcripts and potentially add more insight into what takes place in the minds of the team members. The disadvantage in this type of software is that in order for the software to be able to discern between the dimensions, especially the ones associated with emotions, the text needs to contain significant linguistic contrast and explicit wording that is typical to these dimensions.

Biometric and proxemics comprise a variety of indicators that have their own limitations and have not been incorporated into any known off-the-shelf commercial software packages. The advantage of these methods is that, unlike text-based indicators, these indicators rely on less explicit cues which may not be expressed through text. The disadvantage lies in the fact that more research may be needed in these areas in order to adjust these indicators to the space flight environment and interaction.

The interviews we conducted with personnel currently involved in observing or monitoring astronauts helped us obtain their perspectives on monitoring team performance during long-duration missions and the feasibility of a potential automatic non-obtrusive monitoring system. Their input suggests that negative reactions to monitoring may occur, and that there is a need to convince the astronauts of the importance and benefits of automatic monitoring.

Together, the literature and evidence we reviewed suggest that unobtrusive monitoring of space flight personnel is likely to be a valuable tool for assessing team functioning in future missions. Similar to results from research on electronic monitoring in industrial environments, it is important to have “buy-in” from the personnel who are affected by such monitoring. Certainly, keeping monitoring unobtrusive will help with this process, but the uses and outcomes of monitoring are important dimensions influencing acceptance as well. Several research gaps must be filled in our understanding of what indicators to collect and what analyses to apply before prototype systems can be developed that will provide data about team effectiveness.

17. FUTURE STEPS AND RESEARCH RECOMMENDATIONS

The literature review and operational assessment presented in this document described some of the directions one might pursue in order to design and eventually create systems that would enable monitoring team outcomes based on various indicators. Some of the indicators mentioned above will require more research and adaptation than others. When considering these options, note that the composition of the “team” under study might be

considered not only as the space flight crew itself, but also as a larger collective that includes the ground control personnel. This is a particularly important consideration given data from the interviews: ground crew members may be able to contribute perceptual or objective “criterion” data about the status and performance of space flight teams. These or other criterion data will be required to assess the usefulness of the various indicators being examined. Although it was beyond the scope of the current review, it would be valuable for future research to develop a “directory” of available space flight team performance criteria (e.g., time on task, task error rates) for use in validation studies. With the present state of the art, even if we had the means of collecting reliable predictors of team performance, we might have difficulty validating them for lack of systematically collected criteria.

Although biometrics and proxemics are interesting and promising areas, they still require considerable additional research and evaluation before a workable unobtrusive monitoring system could be designed and implemented. Such an effort would probably need to begin with human review and coding of videotape or other data streams in an effort to observe patterns with relevance to team performance. Given a preliminary understanding of those patterns, both hardware and software prototypes would be needed in order to perform a proof of concept for measures gathered from biometric traces or proxemic cues. A combination of several types of data such as facial expressions, gestures, speech, skin temperature, and proxemic cues might be able to provide a relatively complete picture of team interactions and functioning.

These types of data could preferably be gathered by a small, wearable data collection device, such as the Actiwatch. For example, if the Actiwatch or a similar product could be enhanced to record the proximity between two individuals, measure skin temperature, and record the speech timing of each team member, this might provide a rich source of data for later analysis. A device similar to LOGOPORT (Krüger & Vollrath, 1996), that analyzes speech patterns, could be either part of the device itself or could collect recordings for later analysis. Much of this work could be piloted in analog environments with the beneficial side effect that large data sets of sensor data might serve as a resource for additional research.

Compared with biometrics and proxemics, textual analysis is a more mature and established research area. Notably, open source and commercial software packages are readily available and capable of performing semi-automatic analysis on large text corpora. The disadvantage of textual analysis software packages is that they are currently not adjusted to the requirements of extracting team mental models and often require setup and interpretation by a human operator.

To perform further research and further develop existing text analysis software, there is a crucial need to obtain large corpora of actual communication data such as transcripts of

communication among team members, discussions with ground control personnel, mission logs, and astronauts' personal journals. Thus, an important emerging research need is the collection, transcription, and annotation of “natural” texts that spring from interactions among team members either on the International Space Station or in one or more of the analog environments (e.g., NASA Extreme Environment Mission Operations) where research is conducted. In fact, an initial step in this area would be to conduct a review and feasibility analysis of the various possible sources of text throughout the space flight research ecosystem (including operations in all of the analog environments as well as archival recordings of earlier missions). One valuable goal of such a review would be the development of plans for a data repository, where reusable data, scrubbed to varying levels of anonymity, would become available for use in subsequent research projects.

Mental model elicitation techniques that are currently performed manually by a human operator will need to be translated into software modules or algorithms that will be capable of automatic analysis. If machine learning will be employed, a corpus developed from transcripts and communication texts would be used to train the algorithms to deal with the type of text that represents the space flight domain and terminology. To this end, it will be necessary to develop new dictionaries for use with tools such as LIWC and new workflows for products such as PolyAnalyst. Once the software is enhanced or implemented, the predictive power of communication-related indicators must be confirmed by testing the software, together with a new text corpus and the relevant criterion data, preferably from a current operational environment or analog.

As previously described, it will be essential to involve space flight personnel in the processes of designing, evaluating, and deploying any future monitoring tools. As soon as a promising area of investigation or a candidate technology is selected, further efforts to obtain reaction data from subject matter experts (e.g., space flight personnel who have recently completed one or more missions) using mock-ups of systems and results will help to ensure that a subsequent validation effort or other deployment of a working system will proceed smoothly. Judging by the insights provided from interviewees, the organizational and cultural issues existing among managers, space flight trainees, ground crews, and others may provide substantial barriers to successful implementation, even of a technically sophisticated and effective monitoring system. Thus, rather than focusing exclusively on developing the operational capabilities of a technical solution for unobtrusive team monitoring, we suggest a parallel and simultaneous focus on the “contextual” issues that may enhance or inhibit the successful deployment of tools that can predict space flight team effectiveness. Understanding how to overcome the organizational culture barriers to the deployment of an unobtrusive monitoring system may in the end have equal importance with the technological quality of the system.

18. REFERENCES

- Ajzen, I. (1991). The theory of planned behavior. *Organizational behavior and human decision processes*, 50(2), 179–211.
- Alder, G., Schminke, M., Noel, T., & Kuenzi, M. (2008). Employee Reactions to Internet Monitoring: The Moderating Role of Ethical Orientation. *Journal of Business Ethics*, 80(3), 481-498.
- Alder, G. S. (2001). Employee reactions to electronic performance monitoring: A consequence of organizational culture. *The Journal of High Technology Management Research*, 12(2), 323-342.
- Alder, G. S. (2007). Examining the relationship between feedback and performance in a monitored environment: A clarification and extension of feedback intervention theory. *The Journal of High Technology Management Research*, 17(2), 157-174.
- Alder, G. S., & Ambrose, M. L. (2005). An examination of the effect of computerized performance monitoring feedback on monitoring fairness, performance, and satisfaction. *Organizational Behavior and Human Decision Processes*, 97(2), 161-177.
- Ananyan, S., Kasprzycki, R., & Kollepara, V. (2004). *Application of PolyAnalyst to Flight safety Data at Southwest Airlines*. Megaputer Intelligence.
- André, J., & Funk, P. (2005). A Case Based Approach Using Behavioural Biometrics to Determine a User's Stress Level. In *Workshop proceedings of the 6th International Conference on Case Based Reasoning*. Chicago, editor (s): Isabelle Bichindaritz, Cindy Marling (p. 9).
- Attewell, P. (1987). Big brother and the sweatshop: Computer surveillance in the automated office. *Sociological Theory*, 5, 87–99.
- Barber, A. E., & Roehling, M. V. (1993). Job postings and the decision to interview: A verbal protocol analysis. *Journal of Applied Psychology*, 78, 845–845.
- Bessone, L., Coffey, E., Filippova, N., Greenberg, E., Inoue, N., Gittens, M., Mukai, C., et al. (2008). *International Space Station Human Behavior & Performance Competency Model - Volume II*.
- Bradley, M. M., & Lang, P. J. (1999). *Affective norms for English words (ANEW): Stimuli, instruction manual and affective ratings*. (Technical report). The Center for Research in Psychophysiology, University of Florida.
- Busso, C., Deng, Z., Yildirim, S., Bulut, M., Lee, C. M., Kazemzadeh, A., Lee, S., et al. (2004). Analysis of emotion recognition using facial expressions, speech and multimodal information. In *Proceedings of the 6th international conference on Multimodal interfaces* (pp. 205–211).
- Carley, K. M. (1997). Extracting Team Mental Models through Textual Analysis. *Journal of Organizational Behavior*, 18, 533-558.
- Castellano, G., Kessous, L., & Caridakis, G. (2008). Emotion Recognition through Multiple Modalities: Face, Body Gesture, Speech. In *Affect and Emotion in Human-Computer*

Interaction (pp. 92-103).

- Cavaleri, S., & Serman, J. D. (1997). Towards evaluation of systems thinking interventions: a case study. *System Dynamics Review*, *13*(2), 171–186.
- Chen, J. V., & Ross, W. H. (2007). Individual differences and electronic monitoring at work. *Information, Communication and Society*, *10*, 488-505.
- Chen, Y. (1996). An experimental study of the influences of cognition of interpersonal problems and problem-solving. *Psychological Science China*, *19*, 282–286.
- Cooke, N. J., Neville, K. J., & Rowe, A. L. (1996). Procedural network representations of sequential data. *Human-Computer Interaction*, *11*(1), 29–68.
- Daniels, K., de Chernatony, L., & Johnson, G. (1995). Validating a method for mapping managers' mental models of competitive industry structures. *Human Relations*, *48*(9), 975.
- Davidson, R., & Henderson, R. (2000). Electronic Performance Monitoring: A Laboratory Investigation of the Influence of Monitoring and Difficulty on Task Performance, Mood State, and Self-Reported Stress Levels. *Journal of Applied Social Psychology*, *30*(5), 906-920.
- Deng, L., & Huang, X. (2004). Challenges in adopting speech recognition. *Communications of the ACM*, *47*(1), 69-75.
- D'Urso, S. C. (2006). Who's watching us at work? Toward a structural-perceptual model of electronic monitoring and surveillance in organizations. *Communication Theory*, *16*(3), 281.
- Edwards, B. D., Day, E. A., Arthur, W., & Bell, S. T. (2006). Relationships among team ability composition, team mental models, and team performance. *Journal of Applied Psychology*, *91*(3), 727-736.
- Flint, D., Haley, L., & McNally, J. (2008). Development of a Scale to Measure Reactions to Electronic Monitoring. *The Business Review, Cambridge*, *11*(2), 131.
- Foltz, P. W., Martin, M. J., Abdelali, A., Rosenstein, M. B., & Oberbreckling, R. J. (2006). Automated team discourse modeling: Test of performance and generalization. In *Proceedings of the 28th Annual Cognitive Science Conference* (pp. 1317–1322).
- Francisco, V., & Gervás, P. (2006). Exploring the compositionality of emotions in text: Word emotions, sentence emotions and automated tagging. In *Proceedings of the AAAI-06 Workshop on Computational Aesthetics: AI Approaches to Beauty and Happiness, Boston*.
- Frantova, E., & Bergler, S. (2009). Automatic Emotion Annotation of Dream Diaries. Presented at the K-CAP Workshop on Analyzing Social Media to Represent Collective Knowledge.
- Gentner, D., & Stevens, A. L. (Eds.). (1983). *Mental models*. Routledge.
- Goomas, D. T. (2007). Electronic performance self-monitoring and engineered labor standards for “man-up” drivers in a distribution center. *Journal of Business and Psychology*, *21*(4), 541–558.

- Gorman, J. C., Foltz, P. W., Kiekel, P. A., Martin, M. J., & Cooke, N. J. (2003). Evaluation of Latent Semantic Analysis-Based Measures of Team Communications Content. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 47, 424-428.
- Guinn, C., & Hubal, R. (2003). Extracting emotional information from the text of spoken dialog. In *Proceedings of the 9th International Conference on User Modeling*.
- Guzzo, R. (1995). Introduction: At the intersection of team effectiveness and decision making. *Team effectiveness and decision making in organizations*, (p. 1-8), New York: Sage Publications.
- Hall, E. T. (1963). A System for the Notation of Proxemic Behavior. *American Anthropologist*, New Series, 65(5), 1003-1026.
- Hall, E. T. (1973). *Handbook for proxemic research*. Washington, DC: Society for the Anthropology of Visual Communication.
- Hare, A. (2003). Roles, relationships, and groups in organizations: Some conclusions and recommendations. *Small Group Research*, 34(2), 123.
- Harrigan, J., Rosenthal, R., & Scherer, K. (2008). *New Handbook of Methods in Nonverbal Behavior Research* (1st ed.). Oxford University Press, USA.
- Hassebrock, F., & Prietula, M. J. (1992). A protocol-based coding scheme for the analysis of medical reasoning. *International Journal of Man-Machine Studies*, 37(5), 613-652.
- Hegarty, M. (1991). Knowledge and processes in mechanical problem solving. In R. J. Sternberg & P. A. Frensch (Eds.), *Complex problem solving* (pp. 253-285). Routledge.
- Khan, M. M., Ward, R. D., & Ingleby, M. (2009). Classifying pretended and evoked facial expressions of positive and negative affective states using infrared measurement of skin temperature. *ACM Transactions on Applied Perception (TAP)*, 6(1), 6.
- Kiekel, P. A., Cooke, N. J., Foltz, P. W., Gorman, J. C., & Martin, M. J. (2002). Some Promising Results of Communication-Based Automatic Measures of Team Cognition. *Human Factors and Ergonomics Society Annual Meeting Proceedings*, 46, 298-302.
- Kiker, D., & Kiker, M. (2008). A Quantitative Review of Organizational Outcomes Related to Electronic Performance Monitoring. *The Business Review, Cambridge*, 11(1), 295.
- Klimoski, R., & Mohammed, S. (1994). Team Mental Model: Construct or Metaphor? *Journal of Management*, 20(2), 403-437.
- Kraiger, K., & Wenzel, L. H. (1997). Conceptual development and empirical evaluation of measures of shared mental models as indicators of team effectiveness. *Team performance assessment and measurement: Theory, methods, and applications*, 63-84.
- Krüger, H. P., & Vollrath, M. (1996). Temporal analysis of speech patterns in the real world using the LOGOPORT. In *Ambulatory Assessment. Computer-Assisted Psychological and Psychophysiological Methods in Monitoring and Field Studies* (pp. 103-113). Seattle: Hogrefe & Huber.
- Landauer, T. K., Foltz, P. W., & Laham, D. (1998). An introduction to latent semantic analysis. *Discourse processes*, 25, 259-284.

- Langan-Fox, J., Code, S., & Langfield-Smith, K. (2000). Team Mental Models: Techniques, Methods, and Analytic Approaches. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 42(2), 242-271. doi:10.1518/001872000779656534
- Langan-Fox, J., & Tan, P. (1997). Images of a culture in transition: Personal constructs of organizational stability and change. *Journal of Occupational and Organizational Psychology*, 70, 273–294.
- Langfield-Smith, K. (1992). Exploring the need for a shared cognitive map. *Journal of management studies*, 29(3), 349–368.
- Lanz, O., Brunelli, R., Chippendale, P., Voit, M., & Stiefelhagen, R. (2009). Extracting Interaction Cues: Focus of Attention, Body Pose, and Gestures. In A. Waibel (Ed.), *Computers in the Human Interaction Loop* (pp. 87-93).
- Lim, B. C., & Klein, K. J. (2006). Team mental models and team performance: A field study of the effects of team mental model similarity and accuracy. *Journal of Organizational Behavior*, 27(4), 403-408.
- Markiczy, L., & Goldberg, J. (1995). A method for eliciting and comparing causal maps. *Journal of Management*, 21(2), 305-333.
- Marks, M. A., Mathieu, J. E., & Zaccaro, S. J. (2001). A Temporally Based Framework and Taxonomy of Team Processes. *The Academy of Management Review*, 26(3), 356-376.
- Martin, M. J., & Foltz, P. W. (2004). Automated team discourse annotation and performance prediction using LSA. In *Proceedings of HLT-NAACL 2004: Short Papers on XX* (pp. 97-100). Boston, Massachusetts: Association for Computational Linguistics.
- Mathieu, J. E., Heffner, T. S., Goodwin, G. F., Salas, E., & Cannon-Bowers, J. A. (2000). The influence of shared mental models on team process and performance. *Journal of Applied Psychology*, 85(2), 273-283.
- Mishne, G. (2005). Experiments with mood classification in blog posts. In *Proceedings of ACM SIGIR 2005 Workshop on Stylistic Analysis of Text for Information Access*.
- Mohammed, S., Klimoski, R., & Rentsch, J. R. (2000). The Measurement of Team Mental Models: We Have No Shared Schema. *Organizational Research Methods*, 3(2), 123-165.
- Moran, S., & Nakata, K. (2009). Ubiquitous Monitoring and Human Behaviour in Intelligent Pervasive Spaces. In *Computational Science and Engineering, IEEE International Conference on* (Vol. 4, pp. 1082-1087). Los Alamitos, CA, USA: IEEE Computer Society.
- Nakasone, A., Prendinger, H., & Ishizuka, M. (2005). Emotion recognition from electromyography and skin conductance. In *Proc. of the 5th International Workshop on Biosignal Interpretation* (pp. 219–222).
- NASA - Mission Operations Directorate Space flight Training Management Office. (2009). *International Space Station Astronaut Candidate Training Catalog* (No. JSC-36543). Houston, Texas.
- Panina, D., & Aiello, J. R. (2005). Acceptance of electronic monitoring and its consequences in

- different cultural contexts: A conceptual model. *Journal of International Management*, 11(2), 269-292.
- Park, S., & Catrambone, R. (2007). Social Facilitation Effects of Virtual Humans. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 49(6), 1054-1060.
- Rasmussen, J., & Jensen, A. (1974). Mental procedures in real-life tasks: A case study of electronic trouble shooting. *Ergonomics*, 17(3), 293–307.
- Rouse, W. B., Cannon-Bowers, J. A., & Salas, E. (1992). The role of mental models in team performance in complex systems. *IEEE transactions on systems, man, and cybernetics*, 22(6), 1296–1308.
- Rowe, A. L., & Cooke, N. J. (1995). Measuring mental models: Choosing the right tools for the job. *Human Resource Development Quarterly*, 6(3), 243-255.
- Rubin, V. L., Stanton, J. M., & Liddy, E. D. (2004). Discerning emotions in texts. In *The AAAI Symposium on Exploring Attitude and Affect in Text (AAAI-EAAT)*.
- Schvaneveldt, R. W. (Ed.). (1990). *Pathfinder associative networks: Studies in knowledge organizations*. Ablex Norwood, NJ.
- Sexton, J. B., & Helmreich, R. L. (2003). Using language in the cockpit: Relationships with workload and performance. In R. Dietrich & T. von Meltzer (Eds.), *Communication in high risk environments* (Vol. 12, pp. 57–74). Helmut Buske Verlag.
- Shriberg, E. (2005). Spontaneous speech: How people really talk and why engineers should care. In *Ninth European Conference on Speech Communication and Technology*.
- Smith-Jentsch, K. A., Campbell, G. E., Milanovich, D. M., & Reynolds, A. M. (2001). Measuring Teamwork Mental Models to Support Training Needs Assessment, Development, and Evaluation: Two Empirical Studies. *Journal of Organizational Behavior*, 22(2), 179-194.
- Staggers, N., & Norcio, A. F. (1993). Mental models: concepts for human-computer interaction research. *International Journal of Man-machine studies*, 38(4), 587–605.
- Stanton, J. M., & Julian, A. L. (2002). The impact of electronic monitoring on quality and quantity of performance. *Computers in Human Behavior*, 18(1), 85-101.
- Stanton, J. M., & Stam, K. R. (2006). *The Visible Employee: Using Workplace Monitoring and Surveillance to Protect Information Assets-Without Compromising Employee Privacy or Trust*. Medford, NJ: Information Today, Inc.
- Stout, R. J., Cannon-Bowers, J. A., Salas, E., & Milanovich, D. M. (1999). Planning, Shared Mental Models, and Coordinated Performance: An Empirical Link Is Established. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, 41(1), 61-71.
- Strapparava, C., & Mihalcea, R. (2008). Learning to identify emotions in text. In *Proceedings of the 2008 ACM symposium on Applied computing* (pp. 1556-1560). Fortaleza, Ceara, Brazil: ACM.
- Vandenplas-Holper, H. C. (1996). Intra-individual and inter-individual cognitive conflict, related variables, and relations with cognitive development. *Swiss Journal of Psychology*, 55,

161–175.

- Ververidis, D., & Kotropoulos, C. (2006). Emotional speech recognition: Resources, features, and methods. *Speech Communication, 48*(9), 1162–1181.
- Vizer, L. M., Zhou, L., & Sears, A. (2009). Automated stress detection using keystroke and linguistic features: An exploratory study. *International Journal of Human-Computer Studies, 67*(10), 870-886.
- Watson, A. M. (2008, March 30). *Electronic Monitoring Relevance and Justification: Implications for Procedural Justice and Satisfaction*. North Carolina State University.
- Weitz, S. (1972). Attitude, voice, and behavior: A repressed affect model of interracial interaction. *Journal of Personality and Social Psychology, 24*(1), 14-21. doi:10.1037/h0033383
- Wells, D. L., Moorman, R. H., & Werner, J. M. (2007). The impact of the perceived purpose of electronic performance monitoring on an array of attitudinal variables. *Human Resource Development Quarterly, 18*(1), 121.
- Winitzky, N., Kauchak, D., & Kelly, M. (1994). Measuring teachers' structural knowledge. *Teaching and Teacher Education, 10*(2), 125–139.
- Zhe, X., & Boucouvalas, A. C. (2002). Text-to-emotion engine for real time internet communication. In *Proceedings of International Symposium on Communication Systems, Networks and DSPs* (pp. 164–168).

REPORT DOCUMENTATION PAGE			Form Approved OMB No. 0704-0188	
Public reporting burden for this collection of information is estimated to average 1 hour per response, including the time for reviewing instructions, searching existing data sources, gathering and maintaining the data needed, and completing and reviewing the collection of information. Send comments regarding this burden estimate or any other aspect of this collection of information, including suggestions for reducing this burden, to Washington Headquarters Services, Directorate for Information Operations and Reports, 1215 Jefferson Davis Highway, Suite 1204, Arlington, VA 22202-4302, and to the Office of Management and Budget, Paperwork Reduction Project (0704-0188), Washington, DC 20503.				
1. AGENCY USE ONLY (Leave Blank)	2. REPORT DATE June 2011	3. REPORT TYPE AND DATES COVERED NASA Technical Memorandum		
4. TITLE AND SUBTITLE Unobtrusive Monitoring of Spaceflight Team Functioning			5. FUNDING NUMBERS	
6. AUTHOR(S) Veronica Maidel, M.S., Jeffrey M. Stanton, Ph.D., Syracuse University, Syracuse, NY Kathryn E. Keeton, Ph.D., NASA Johnson Space Center, Houston, TX				
7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES) Lyndon B. Johnson Space Center Houston, Texas 77058			8. PERFORMING ORGANIZATION REPORT NUMBERS S-1100	
9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES) National Aeronautics and Space Administration Washington, DC 20546-0001			10. SPONSORING/MONITORING AGENCY REPORT NUMBER TM-2011-216153	
11. SUPPLEMENTARY NOTES				
12a. DISTRIBUTION/AVAILABILITY STATEMENT Available from the NASA Center for AeroSpace Information (CASI) 7121 Standard Hanover, MD 21076-1320 Category: 53			12b. DISTRIBUTION CODE	
13. ABSTRACT (Maximum 200 words) This document contains a literature review suggesting that research on industrial performance monitoring has limited value in assessing, understanding, and predicting team functioning in the context of space flight missions. The review indicates that a more relevant area of research explores the effectiveness of teams and how team effectiveness may be predicted through the elicitation of individual and team mental models. The "mental models" referred to in this literature typically reflect a shared operational understanding of a mission setting such as the cockpit controls and navigational indicators on a flight deck. In principle, however, mental models also exist pertaining to the status of interpersonal relations on a team, collective beliefs about leadership, success in coordination, and other aspects of team behavior and cognition. Conclusions from this work suggest that unobtrusive monitoring of space flight personnel is likely to be a valuable future tool for assessing team functioning, but that several research gaps must be filled before prototype systems can be developed for this purpose.				
14. SUBJECT TERMS astronaut performance, biometrics, cognition, monitoring			15. NUMBER OF PAGES 76	16. PRICE CODE
17. SECURITY CLASSIFICATION OF REPORT Unclassified	18. SECURITY CLASSIFICATION OF THIS PAGE Unclassified	19. SECURITY CLASSIFICATION OF ABSTRACT Unclassified	20. LIMITATION OF ABSTRACT Unlimited	
