2006-02

Assisting People to Become Independent Learners in the Analysis of Intelligence

Pirolli, Peter

Assisting People to Become Independent Learners in the Analysis of Intelligence

Final Technical Report

Office of Naval Research
Contract N00014-02-C-0203
CDRL A002

Author: Peter Pirolli

Palo Alto Research Center, Inc.
3333 Coyote Hill Road
Palo Alto, CA. 94304

REPORT DOCUMENTATION PAGE

Form Approved
OMB No. 0704-0188

1. REPORT DATE *(DD-MM-YYYY)*	2. REPORT TYPE	3. DATES COVERED *(From - To)*
02-28-2006	Final Report	01-31-2002 to 12-31-2005

4. TITLE AND SUBTITLE
Assisting People to Become Independent Learners in the Analysis of Intelligence: Final Technical report (CDRL A002)

5a. CONTRACT NUMBER
N00014-02-C-0203

5b. GRANT NUMBER

5c. PROGRAM ELEMENT NUMBER

6. AUTHOR(S)
Pirolli, Peter L.

5d. PROJECT NUMBER

5e. TASK NUMBER

5f. WORK UNIT NUMBER

7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
Palo Alto Research Center, Inc
3333 Coyote Hill Rd
Palo Alto, CA 94304

8. PERFORMING ORGANIZATION REPORT NUMBER
ONR 2002 Final Report

9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)
Office of Naval Research
875 North Randolph Street, Suite 1425
Code 34 and Code 254
Arlington, CA 22203-1995

10. SPONSOR/MONITOR'S ACRONYM(S)
ONR

11. SPONSORING/MONITORING AGENCY REPORT NUMBER

12. DISTRIBUTION AVAILABILITY STATEMENT
Approved for public release; distribution unlimited

13. SUPPLEMENTARY NOTES

14. ABSTRACT

The purpose of this project was to conduct applied research with exemplary technology to support post-graduate instruction in intelligence analysis. The first phase of research used Cognitive Task Analysis (CTA) to understand the nature of subject matter expertise for this domain, as well as leverage points for technology support. Results from the CTA and advice from intelligence analysis instructors at the Naval Postgraduate School lead us to focus on the development of a collaborative computer tool (CACHE) to support a method called the Analysis of Competing Hypotheses (ACH). We first evaluated a non-collaborative version of an ACH tool in an NPS intelligence classroom setting, followed by an evaluation of the collaborative tool, CACHE at NPS. These evaluations, along with similar studies conducted in coordination with NIST and MITRE, suggested that ACH and CACHE can support intelligence activities and mitigate confirmation bias. However, collaborative analysis has a number of trade-offs: it incurs overhead costs, and can mitigate or exacerbate confirmation bias, depending on the mixture of predisposing biases of collaborators.

15. SUBJECT TERMS
Intelligence analysis

16. SECURITY CLASSIFICATION OF:			17. LIMITATION OF ABSTRACT	18. NUMBER OF PAGES	19a. NAME OF RESPONSIBLE PERSON
a. REPORT	b. ABSTRACT	c. THIS PAGE			Peter Pirolli
unclassified	unclassified	unclassified	unclassified	99	19b. TELEPONE NUMBER (*Include area code*) 650-812-4483

OVERVIEW

The purpose of this project (*Assisting People to Become Independent Learners in the Analysis of Intelligence*) was to conduct applied research with exemplary technology to support post-graduate instruction in intelligence analysis. Part of this effort was aimed at building up empirical studies of analyst knowledge, reasoning, performance, learning, etc, involved in intelligence analysis. This first phase of research used Cognitive Task Analysis (CTA) to understand the nature of subject matter expertise for this domain, as well as leverage points for technology support. Results from the CTA and advice from intelligence analysis instructors at the Naval Postgraduate School lead us to focus on the development of a collaborative computer tool (CACHE) to support a method developed by Heuer (1999) called the Analysis of Competing Hypotheses (ACH). We first evaluated a non-collaborative version of an ACH tool in an NPS intelligence classroom setting, followed by an evaluation of the collaborative tool, CACHE at NPS. These evaluations, along with similar studies conducted in coordination with NIST and MITRE, suggested that ACH and CACHE can support intelligence activities and mitigate confirmation bias. However, collaborative analysis has a number of trade-offs: it incurs overhead costs, and can mitigate or exacerbate confirmation bias, depending on the mixture of predisposing biases of collaborators.

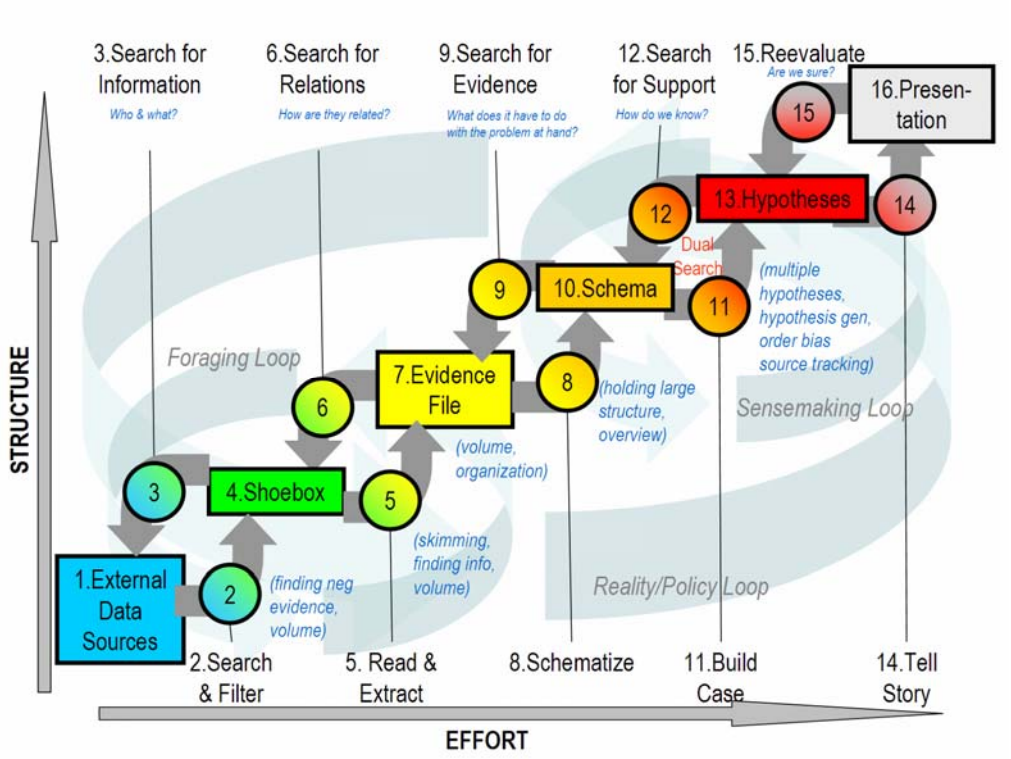

Figure 1. Notional model of the intelligence analyst process.

Intelligence analysis is an example of a class of tasks we have come to call *sensemaking* tasks. Such tasks involve finding and collecting information from large information collections, organizing

and understanding that information, and producing some product, such as a briefing or actionable decision. Examples of such tasks include understanding a health problem in order to make a medical decision, forecasting the weather, or deciding which laptop to buy. Figure 1 represents our notional understanding of the analyst's process. The rectangular boxes represent an approximate data flow we have seen across several analysts. The circles represent the process flow. The processes and data are arranged by degree of effort and degree of structure. This is a process with lots of back loops and seems to have one set of activities that cycle around finding information and another that cycles around making sense of the information, with plenty of interaction between these. This process diagram summarizes how it is that an analyst comes up with *novel information*. The overall process is organized into two major loops of activities: (1) a *foraging loop* (Figure 1) that involves processes aimed at seeking information, searching and filtering it, and reading and extracting information, and (2) a *sense making loop* that involves iterative development of a mental model (a conceptualization) that best fits the evidence. Information processing can be driven by *bottom-up* processes (from data to theory) or *top-down* (from theory to data in an opportunistic mix.

Section 1 of this report provides an account of our CTA of the process depicted in Figure 1. That analysis, in combination with other CTAs (Pirolli *et al*., 2004), suggested that we focus on (a) *foraging*; improving the amount of useful data processed per unit time and (b) *sensemaking*; mitigation of confirmation bias. With ACH and CACHE the aim was to achieve these effects by reducing the cost structure of interaction, improving the coverage of data (the total set collectively attended) through collaboration, and technology that supported attention to alternative hypotheses and a process of disconfirmation rather than confirmation.

Section 2 reports on our evaluation of a computer-based tool to support the ACH method. The basic ACH method involves using a Hypothesis X Evidence matrix, identifying possible hypotheses, listing evidence for/against each hypothesis, and attempting to disprove hypotheses as opposed to confirming hypothesis. The basic claims about the ACH method (Heuer, 1999) are that it promotes the generation of a fuller set of alternative hypotheses, alternative hypotheses receive more equitable distribution of attention, attention is focused on evidence with the greatest diagnostic value, and confirmation bias is reduced. Cheikes *et al*. (2004) performed a study of a non-computerized version of the ACH method and found that, although all subjects showed evidence of confirmation bias, users of ACH showed less distortions of evidence than non-ACH users. These effects were stronger for non-expert analysts than experts. An unpublished study by Emile Morse and Jean Scholtz at NIST had analysts work cases with and without the ACH tool developed at PARC. Among other findings, the NIST study found that analysts were confident that ACH would help improve the thoroughness of analysis, the method was easy to learn and use, and they were inclined to us the method in future work. The PARC evaluation of the ACH tool showed that the computer version of ACH did not differ from doing ACH by hand in terms of amount of evidence considered or hypotheses generated.

Section 3 describes an evaluation of a follow-on system to ACH called CACHE. Using feedback on ACH, we developed a Collaborative ACH Environment (CACHE) designed to support collaboration (synchronous or asynchronous) among communities or groups of analysts. CACHE is designed to support the sharing of evidence and of ACH analysis matrices. The CACHE architecture is a Web-based, client-server architecture. The user interface can be run through any Web browser, which connects to a backend that supports evidence integration, natural language processing, and semantic search. Generally, there is the intuition that one can achieve better intelligence analysis through collaboration because: (1) Like over-the-horizon radar, an individual analyst may receive information otherwise unseen because of the information flowing to him or her from a social network of collaborators. (2) Collectively, by arranging the spotlights of attention of individual analysts to insure maximum, exhaustive coverage of the evidence and hypotheses, one can bring to light some crucial data or insight that might otherwise be missed. (3) Coordinated teams of experts may be assembled in order to exploit years of specialized skill and relevant knowledge about background and precedents, and to deliberately confront problems whose solutions require breakthroughs. (4) Diversity of viewpoints can be brought to bear to provide mutually corrective forces to overcome the cognitive heuristics and biases that often create blindness to impending threats. Each of these beneficial effects is crucial to overcoming the most frequent impediments to intelligence analysts. However, there are also costs to collaboration. As reviewed in Section 3, face-to-face interaction frequently yields biased information foraging because there is a tendency to focus on evidence and hypotheses held in common. The CACHE experiment in Section 3 was motivated by prior research indicating that computer support can mitigate this “common ground” effect, and careful construction of groups to insure diversity could mitigate confirmation bias. The results of the CACHE evaluation were mixed. One the one hand, careful construction of diverse groups mitigated confirmation bias as compared to collaborations among groups with homogenous biases, However, diverse groups and solitary users were not significantly different in their reduction in confirmation bias. So, CACHE appears to mitigate confirmation bias in solitary users and diverse groups, but it appears that the overhead of collaboration in CACHE was cancelling out any additional benefits of collaboration over solitary work. However, as this was the first evaluation of CACHE, usability questions revealed a large number of potential improvements that could be made to CACHE to attenuate the collaboration costs in future versions of the system.

REFERENCES

Cheikes, B.A., brown, M.J., Lehrner, P.E.,, and Alderman, L. (2004). *Confirmation bias in complex analyses.* (Tech. Rep. No. MTR 04B0000017). Bedford, MA: MITRE.

Heuer, R.J. (1999). *Psychology of intelligence analysis.* Washington, D.C.: Center for the Study of Intelligence.

Pirolli, P., Lee, T., and Card, S.K. (2004). Levererage points for analyst technology identified through cognitive task analysis (UIR Tech. Rep.) Palo Alto, CA: Palo Alto Research Center.

Section 1
What Makes Intelligence Analysis Difficult?
A Cognitive Task Analysis of Intelligence Analysts

Susan G. Hutchins
Naval Postgraduate School
Information Sciences Department
589 Dyer Road, Code IS/Hs
Monterey, CA 93943-5000
shutchins@nps.edu

Peter L. Pirolli and Stuart K. Card
Palo Alto Research Center
User Interface Research Group
3333 Coyote Hill Road
Palo Alto, CA 94304
pirolli@parc.com
card@parc.com

ABSTRACT

Intelligence analysts engage in information seeking, evaluation, prediction, and reporting behavior in an extremely information-intensive work environment. A Cognitive Task Analysis (CTA) was conducted on intelligence analysts to capture data that will provide input to support development of a computational model of the analyst's processes and analytic strategies. A hybrid method was used to conduct the CTA, including a modified version of the critical decision method. Participants were asked to describe an example of a critical analysis assignment where they had to collect, analyze, and produce a report on intelligence of a strategic nature. Procedures used to conduct the CTA are described in this chapter along with initial results. Several factors contribute to making the analyst's task challenging: (*i*) time pressure, (*ii*) a high cognitive workload, and (*iii*) difficult human judgments. Human judgments are involved in considering the plausibility of information, deciding what information to trust, and determining how much weight to place on specific pieces of data. Intelligence analysis involves a complex process of assessing the reliability of information from a wide variety of sources and combining seemingly unrelated events. This problem is challenging because it involves aspects of data mining, data correlation and human judgment.

INTRODUCTION

In this chapter we describe research involving a Cognitive Task Analysis (CTA) with intelligence analysts, in line with one of the themes of this book, namely, strategies used by experts who are confronted with tough scenarios and unusual tasks. We present what we have learned regarding how experienced practitioners deal with the extremely challenging task of intelligence analysis by summarizing a set of ten CTA interviews conducted with intelligence analysts to identify leverage points for the development of new technologies.

The challenges facing practitioners in the modern world where expertise gets "stretched" by dynamics and uncertainty, a second them for this book, also characterize the problems experienced by intelligence analysts. Part of the effort reported in this chapter is aimed at building up an empirical psychological science of analyst knowledge, reasoning, performance, and learning. We expect this will provide a scientific basis for design insights for new analyst technologies. In addition, this psychological research should yield task scenarios and benchmark tasks that can be used in controlled experimental studies and evaluation of emerging analyst technologies.

INTELLIGENCE ANALYSIS

An ability to sort through enormous volumes of data and combine seemingly unrelated events to construct an accurate interpretation of a situation and make predictions about complex, dynamic events represents the hallmark of the intelligence analysts (IA's) job. These volumes of data typically represent an extensive and far-ranging collection of sources, and are represented in many different formats (e.g., written and oral reports, photographs, satellite images, maps, tables of numeric data, to

name a few). As part of this process, the analyst must make difficult judgments to assess the relevance, reliability, and significance of these disparate pieces of information. Intelligence analysis also involves performing complex reasoning processes such as inferential analysis, to determine "the best explanation for uncertain, contradictory and incomplete data" (Patterson, Roth, & Woods, 2001, p. 225).

The nature of the data, the complex judgments and reasoning required, and a sociotechnical environment that is characterized by high workload, time pressure, and high stakes combine to create an extremely challenging problem for the intelligence analyst. High levels of uncertainty are associated with the data, when "deception is the rule." Since the validity of the data is always subject to question, this impacts the cognitive strategies used by analysts (Johnson, 2004). Moreover, the complex problems to be analyzed entail complex reasoning, including abductive[1], deductive[2] , and inductive[3] reasoning. Finally, high stakes are associated with the pressure not to miss anything and to provide timely, actionable analysis. Potentially high consequences for failure — where analysis products have a significant impact on policy — also contribute to make the task challenging as decisionmakers, senior policy makers, and military leaders use the products of analysis to make high-stakes decisions involving national security.

A number of reports have emerged that provide normative or prescriptive views on intelligence analysis. There have been very few that provide empirical, descriptive

---

1 Abductive reasoning is used to determine the best explanation (Josephson & Josephson, 1994) where if the match between data and an explanation is more plausible than any other explanation it is accepted as the likely explanation (Klein, this volume).

2 Deductive reasoning involves deriving a conclusion by logical deduction; inference in which the conclusion follows the premises.

3 Inductive reasoning employs logical induction where the conclusion, though supported by the premises, does not follow from them necessarily.

studies of intelligence analysis. It is likely that there are many CTA studies of intelligence analysis that will never become part of the public literature because of the classified nature of the work involved. Despite the spottiness of available literature, what does exist reveals that intelligence analysis is a widely variegated task domain. This means that it is important to be careful in making generalizations from any circumscribed types of intelligence tasks or types of analysts. It is equally important not to be daunted by the vastness of the domain, and to start the investigative venture somewhere.

Intelligence analysis is commonly described as a highly iterative cycle involving requirements (problem) specification, collection, analysis, production, dissemination, use, and feedback. It is an event-driven, dynamic process that involves viewing the information from different perspectives in order to examine competing hypotheses and develop an understanding of a complex issue. The critical role of the human is to add "value" to original data by integrating disparate information and providing an interpretation (Krizan, 1999). This integration and interpretation entails difficult, complex judgments to make sense of the information obtained. This "dis-aggregation and synthesis of collected and created evidence includes sorting out the significant from the insignificant, assessing them severally and jointly, and arriving at a conclusion by the exercise of judgment: part induction, part deduction, and part abduction." (Millward, 1993, in Moore, 2003).

Warning-oriented intelligence includes supporting the need for senior policymakers to not be surprised (Bodnar, 2003). Analysts need to "provide detailed enough judgments — with supporting reporting — so that both the warfighter and the policymaker can anticipate the actions of potential adversaries and take timely action to

support US interests" (*ibid.*, p. 6). For example, the analyst needs to make predictions regarding what the adversary has the capability to do and how likely it is that he will act. These predictions need to include what actions can be taken to change, or respond to these actions, and the probable consequences of those actions (*ibid.*).

Table 1 presents an analysis of problem types that Krizan derives from Jones (1995) and course work at the Joint Military Intelligence College. A range of problem types, from simplistic to indeterminate, are explicated by characterizing each level of the problem along several dimensions, such as type of analytic task, analytic method, output, and probability of error.

Table 1. Intelligence Analysis Problem Types (Krizan, 1999).

Taxonomy of Problem Types **Source: Analysis course material, Joint Military Intelligence College, 1991**					
Characteristics	**Problem Types**				
	Simplistic	**Deterministic**	**Moderately Random**	**Severely Random**	**Indeterminate**
What is the question?	Obtain information	How much? How many?	Identify and rank all outcomes	Identify outcomes in unbounded situation	Predict future events/situations
Role of facts	Highest	High	Moderate	Low	Lowest
Role of judgment	Lowest	Low	Moderate	High	Highest
Analytical task	Find information	Find/create formula	Generate all outcomes	Define potential outcomes	Define futures factors
Analytical method	Search sources	Match data to formula	Decision theory; utility analysis	Role playing and gaming	Analyze models and scenarios
Analytical instrument	Matching	Mathematical formula	Influence diagram, utility, probability	Subjective evaluation of outcomes	Use of experts
Analytic output	Fact	Specific value or number	Weighted alternative outcomes	Plausible outcomes	Elaboration on expected future
Probability of error	Lowest	Very low	Dependent on data quality	High to very high	Highest
Follow-up task	None	None	Monitor for change	Repeated testing to determine true state	Exhaustive learning

Figure 1 presents another way of characterizing the domain of intelligence analysis developed by Cooper. Along one axis there are various types of *intelligence*, along a second are different *accounts* (topics), and along a third axis are different types of *products*. The different types of intelligence (or "sources") are functionally organized into:

- *human source intelligence* (HUMINT), which includes field agents, informants, and observers (attaches),
- *imagery intelligence* (IMINT), which includes photo, electro-optical, infrared, radar, and multispectral imagery from sources such as satellites,
- *signals intelligence* (SIGINT), which includes communications, electronic, and telemetry,
- *measurement and signatures intelligence* (MASINT), which includes acoustic and radiation signals,
- *open source intelligence* (OSINT), which includes public documents, newspapers, journals, books, television, radio, and the World Wide Web, and
- all-source intelligence, which involves all of the above.

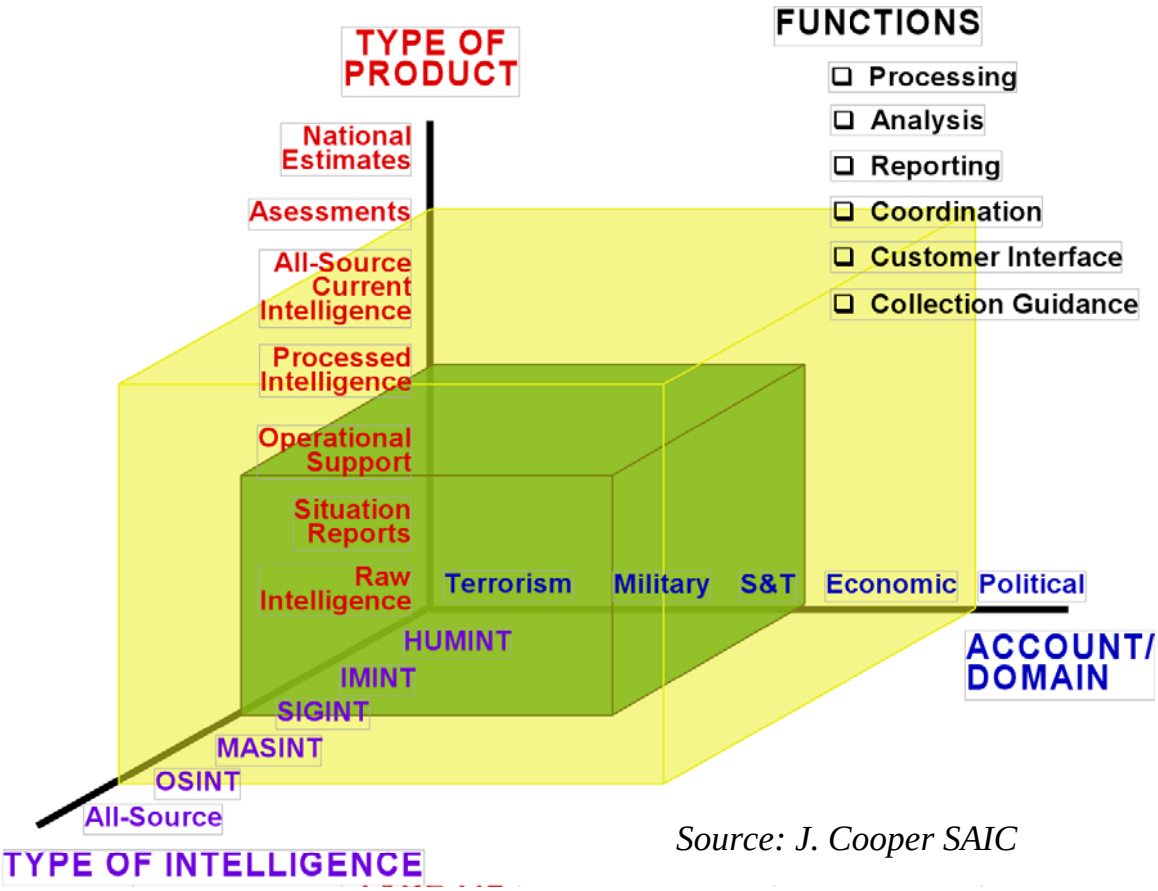

Figure 1. Types of Intelligence, Domains, Functions, and Products.

Domains (or topics) may address terrorism, military, politics, science and technology (S&T), or economics. Product types range from those that are close to the raw data, through those that involve increasing amounts of analysis that may eventually lead to national-level estimates and assessments. As in any hierarchically organized information system, this means that information is filtered and recoded as the analysis process progresses from lower to higher levels.

Techniques to Enhance Processing of Intelligence Data

Recent world events have focused attention on some of the inherent challenges involved in performing intelligence analysis (*viz.*, The 9/11 Commission Report). As a result, increased research is being conducted to develop new training, tools, and

techniques that will enhance the processing of intelligence data. As one example, support and training in the organizing and piecing together aspects of intelligence analysis and decision making has been identified by the Office of the Assistant Secretary of Defense for Networks and Information Integration (OASD/NII) Research Program as an area that is greatly in need of more basic and applied research. One current research thread that seeks to address this need is the Novel Information from Massive Data (NIMD) program where the goal is to develop an "information manager" to assist analysts in dealing with the high volumes and disparate types of data that inundate intelligence analysts. The NIMD research program seeks to develop techniques that "structure data repositories to aid in revealing and interpreting novel contents" and techniques that can accurately model and draw inferences about (1) rare events and (2) sequences of events (widely and sparsely distributed over time).

Connable (2001) asserts that the intelligence process would be well served by enhancing the ability to leverage open sources, particularly since open sources provide the Intelligence Community with between 40-80% of its usable data (Joint Military Intelligence Training Center, 1996). As an example, one of our study participants, who worked on a strategic analysis assignment regarding the question of whether President Estrada, of the Philippines, was going to remain in power or be removed from office, indicated that 80% of the information he needed was found in open-source material. Information foraging theory (Pirolli & Card, 1998; Pirolli & Card, 1999) is being applied in this research on tasks that involve information-intensive work where the approach is to analyze the tasks as an attempt by the user to maximize information gained per unit time. A computational model of the intelligence analysis process will be developed as a result of this CTA research and used to support tool prototyping and testing.

The goals for the research described in this chapter are threefold. One purpose of this first CTA phase is to yield “broad brushstroke” models of analyst knowledge and reasoning at a large grain size of behavioral analysis. A second purpose of this research is to identify leverage points where technical innovations may have the chance to yield dramatic improvements in intelligence analysis. A third purpose of the CTA phase is to guide the development of benchmark tasks, scenarios, resources, corpora, evaluation methods and criteria to shape the iterative design of new analyst technologies. A CTA is typically used to identify the decision requirements, and the knowledge and processing strategies used for proficient task performance. The following section presents a brief description of CTA and describes specific techniques that are representative of CTA methods.

COGNITIVE TASK ANALYSIS

CTA refers to a group of methods that are extensively used in naturalistic decision-making applications. Klein's (2001, p. 173) definition of a CTA is "a method for capturing expertise and making it accessible for training and system design." Klein delineates the following five steps: (1) identifying sources of expertise; (2) assaying the knowledge; (3) extracting the knowledge; (4) codifying the knowledge; and (5) applying the knowledge. System design goals supported by CTA include human-computer interaction design, developing training, tests, models to serve as a foundation for developing an expert system, and analysis of a team's activities to support allocation of responsibilities to individual humans and cooperating computer systems.

Different CTA methods are used for different goals. Our goals for conducting a CTA are twofold. Our first goal is to capture data that will provide input to support development of a computational model of the intelligence analyst's processes and

analytic strategies. Our second goal is to identify leverage points to inform the development of tools to assist analysts in performing the most demanding aspects of their tasks. CTA extends traditional task analysis techniques to produce information regarding the knowledge, cognitive strategies, and goal structures that provide the foundation for task performance (Chipman, Schraagen, & Shalin, 2000). The goal of CTA is to discover the cognitive activities that are required for performing a task in a particular domain to identify opportunities to improve performance by providing improved support of these activities (Potter, Roth, Woods, & Elm, 2000).

Our overall approach for the first phase of this research involves the following steps: review of the intelligence literature, use of semi-structured interviews, followed by the use of structured interviews and review of the results by subject matter experts (SMEs). The second phase for this research, conducted in the summer of 2004, involved developing and comparing several alternative hypotheses based on material presented in a case study. A prototype tool developed to assist the intelligence analyst in comparing alternate hypotheses was introduced and simulated tasks were performed to empirically evaluate the tool's effectiveness. A follow-on study will involve the use of think-aloud protocol analysis while using a more advanced version of this tool. This multiple-phase plan is in line with the approach employed by several successful CTA efforts (Hoffman, et al., 1995; Patterson, Roth, & Woods, 2002). We are using a "balanced suite of methods that allow both the demands of the domain and the knowledge and strategies of domain experts to be captured in a way that enables clear identification of opportunities for improved support." (Potter, et al., 2000, p. 321).

Types of activities that typically require the resource intensive analysis frequently required when conducting a CTA are those domains that are characterized as (*i*)

complex, ill-structured tasks that are difficult to learn, (*ii*) involving complex, dynamic, uncertain, and real-time environments, and (iii) sometimes include multitasking. A CTA is most appropriate when the task requires the use of a large and complex conceptual knowledge base; the use of complex goal/action structures dependent on a variety of triggering conditions, or complex perceptual learning or pattern recognition. Intelligence analysis involves all of these characteristics.

When considering which knowledge elicitation technique is most appropriate, the differential access hypothesis proposes that different methods elicit different types of knowledge (Hoffman, Shadbolt, Burton, & Klein, 1995). Certain techniques are appropriate to "bootstrap" the researcher and generate an initial knowledge base and more structured techniques are more appropriate to validate, refine and extend the knowledge base (*ibid*). A direct mapping should exist between characteristics of the targeted knowledge and the technique/s selected (Cooke, Salas, Cannon-Bowers, & Stout, 2002).

A detailed, accurate cognitive model that delineates the essential procedural and declarative knowledge is necessary to develop effective training procedures and systems (Annett, 2000). This entails building a model that captures the analysts' understanding of the demands of the domain, the knowledge and strategies of domain practitioners, and how existing artifacts influence performance. CTA can be viewed as a problem-solving process where the questions posed to the subject-matter experts, and the data collected, are tailored to produce answers to the research questions, such as training needs and how these training problems might be solved (DuBois & Shalin, 2000). A partial listing of the types of information to be obtained by conducting a CTA includes factors that contribute to making task performance challenging, what strategies

are used and why, what complexities in the domain practitioners respond to, what aspects of performance could use support, concepts for aiding performance, and what technologies can be brought to bear to deal with inherent complexities.

Use of Multiple Techniques

Analysis of a complex cognitive task, such the intelligence analyst's job, often requires the use of multiple techniques. When results from several techniques converge confidence is increased regarding the accuracy of the CTA model (Cooke, 1994; Flach, 2000; Hoffman, et al., 1995; Potter, et al., 2000). Flach (2000) recommends sampling a number of experts and using a variety of interviewing tools to increase the representativeness of the analysis. During the initial bootstrapping phase of this research, several CTA approaches were examined with an eye toward determining which approach would be most productive for our domain of interest. The remainder of this section describes two CTA techniques that were used for the initial phase of this research.

Applied Cognitive Task Analysis Method. Our initial set of interviews drew upon the Applied Cognitive Task Analysis (ACTA) Method (Militello & Hutton, 1998; Militello *et al.*, 1997) and the Critical Decision Method (Hoffman, Coffey, & Ford, in press; Hoffman, Crandall, & Shadbolt, 1998; Klein, Calderwood, & MacGregor, 1989). The ACTA collection of methods was developed explicitly as a streamlined procedure for instructional design and development (Militello et al., 1997) that required minimal training for task analysts. ACTA is a collection of semi-structured interview techniques that yields a general overview of the SMEs' conception of the critical cognitive processes involved in their work, a description of the expertise needed to perform

complex tasks, and SME identification of aspects of these cognitive components that are crucial to expert performance.

The standard ACTA methodology[4] includes the use of three interview protocols and associated tools: (a) the Task Diagram, (b) the Knowledge Audit and (c) the Simulation Overview. The ACTA Method uses interview techniques to elicit information about the tasks performed and provides tools for representing the knowledge produced (Militello & Hutton, 1998). Discovery of the difficult job elements, understanding expert strategies for effective performance, and identification of errors that a novice might make are objectives for using the ACTA method. The focus for researchers using the ACTA method is on interviews where domain practitioners describe critical incidents they have experienced while engaged in their tasks and aspects of the task that made the task difficult.

Our use of the ACTA method produced valuable data for the initial bootstrapping phase of this research where the goal was to learn about the task, the cognitive challenges associated with task performance, and to determine what tasks to focus on during ensuing phases of the CTA research. Products typically produced when using the ACTA method include a Knowledge Audit and a Cognitive Demands Table. After conducting this first group of CTA interviews we opted to use a different method to capture the essence of the IA's job. The IA's task places greater emphasis on deductive and inductive reasoning, looking for patterns of activity, and comparing hypotheses to make judgments about the level of risk present in a particular situation. We felt it was necessary to broaden the scope of the interview probes used with intelligence analysts.

---

[4] Software available from Klein Associates provides rapid training plus interview materials for ACTA.

Critical Decision Method. The Critical Decision Method (CDM) is a semi-structured interview technique developed to obtain information about decisions made by practitioners when performing their tasks. Specific probe questions help experts describe what their task entails. CDM's emphasis on non-routine or difficult incidents produces a rich source of data about the performance of highly skilled personnel (Hoffman, Crandall, & Shadbolt, 1998; Hoffman, Coffey, & Ford, in press; Klein, Calderwood, & MacGregor, 1989). By focusing on critical incidents, the CDM is efficient in uncovering elements of expertise that might not be found in routine incidents and helps to ensure a comprehensive coverage of the subject matter.

Our use of the CDM was tailored to develop domain-specific cognitive probes that elicit information on how analysts obtain and use information, schemas employed to conceptualize the information, how hypotheses are developed to analyze this information, and the types of products that are developed as a result of their analysis. A strength of the CDM is the generation of rich case studies, including information about cues, hypothetical reasoning, strategies, and decision requirements (Klein, et al., Hoffman, Coffey, Carnot, & Novak, 2002). This information can then be used in modeling the reasoning procedures for a specific domain.

In the remainder of this chapter we describe the development and use of an adapted version of the CDM and results derived from use of two CTA methods, ACTA and CDM.

METHOD

Procedures used to conduct the CTA, using ACTA and the CDM, are described in this section as study 1 and study 2, respectively. In the first study we learned about the task, the cognitive challenges associated with task performance, and determined

what tasks to focus on during ensuing phases of the CTA research. In the second study we revised the methodology and used a different group of IAs. Interview probes were developed and used to conduct an adapted version of the CDM where participants were asked to describe a strategic *analysis* problem in lieu of a critical decision problem.

STUDY 1

Participants

Six military intelligence analysts, currently enrolled in a graduate school program at the Naval Postgraduate School (NPS), Monterey, CA, were interviewed for the first study. Participants were contacted via e-mail with the endorsement of their curriculum chair and were asked to volunteer for this study. (No rewards were given for participation.) These U.S. Naval officers (Lieutenant through Lieutenant Commander) were students in the Intelligence Information Management curricula at NPS.

Participants in both studies had an average of ten years experience working as intelligence analysts. Thus, they were considered experts as the literature generally defines an "expert" as an individual who has over ten years experience and "would be recognized as having achieved proficiency in their domain" (Klein, et al., 1989, p. 462).

Materials

Study participants (study 1 and 2) had pen and paper, and a flip chart or white board. After a brief introduction to the study participants were asked to complete a demographic survey.

Procedure

The CTA process for all study participants took place in a small conference room at NPS. Semi-structured interviews were conducted with the first group of interviewees

where intelligence analysts were asked to recall and describe an incident from past job experience.

ACTA. Domain experts were asked to draw a task diagram, to describe critical incidents they had experienced on their job, and identify examples of the challenging aspects of their tasks. They were asked to elucidate why these tasks are challenging, and to describe the cues and strategies that are used by practitioners, and the context of the work. Interviews were scheduled for one and one-half hours at a time that was convenient for each participant. Three interviewers were present for each of the first six interviews. The interviews were tape-recorded and transcribed and the analysis was performed using the transcription and any other materials produced during the interview, e.g., task diagrams.

This first group of intelligence analysts had a variety of assignments in their careers, however the majority of their experience was predominantly focused on performing analysis at the tactical level. (Tactical level analysis refers to analysis of information that will impact mission performance within the local operating area, e.g., of the battle group, and generally within the next 24 hours.) During this bootstrapping phase of our CTA effort, we learned that there are several career paths for intelligence analysts. These career paths can be categorized as either having more of a technology emphasis where the focus is on systems, equipment, and managing the personnel who operate and maintain this equipment or an analytical emphasis where the focus and experience is on performing long-range, or strategic, analysis.

Information gathered during the initial phase served as an advance organizer by providing an overview of the task and helped to identify the cognitively complex elements of the task. The ACTA method produced valuable data for the initial phase of

this research. After analyzing the data from the initial set of interviews, we determined that we needed to broaden the set of interview probes and tailor them for the specific domain of intelligence analysis to uncover the bigger picture of how intelligence analysts approach performing their job. Thus, tailored probes were developed specifically for the domain of intelligence analysis.

Concurrent with the decision to use an adapted version of the CDM was the decision to switch to a different group within the intelligence community, specifically analysts who had experience at the strategic, or national, level.[5] National level intelligence is more concerned with issues such as people in positions of political leadership, and the capabilities of another country. In contrast, at the tactical level, the user of intelligence information may only be concerned about a specific ship that is in a particular area, at a certain time; that is, the information will only be valid for a limited time. Descriptions of experiences at the tactical level did not provide examples of the types of problems or cases that could benefit from the technology envisioned as the ultimate goal for this research.

STUDY 2

Participants

Four military intelligence analysts from the National Security Affairs (NSA) Department were interviewed for the second study. In the NSA curriculum there is a stronger analytical emphasis and the analysts have had experience with analysis assignments at the strategic level. We were fortunate in that this second group of participants was very articulate in describing assignments where they had performed

[5] The term 'strategic analysis' can have several definitions. We are referring to intelligence problems that have implications of strategic importance and those that require more time than is devoted to tactical questions, i.e., analysis tasks that require anywhere from several weeks to many months (or even years) to complete.

analysis of critical topics at the strategic level. Several researchers have noted the issue of encountering problems with inaccessible expert knowledge (Cooke, 1994; Hoffman, Shadbolt, Burton, & Klein, 1995).

Procedure

Structured interviews were conducted with the second group of interviewees where intelligence analysts were asked to recall a strategic analysis problem they had worked on. Participants were asked to describe what they did step-by-step to gather the information and analyze it, and to construct a timeline to illustrate the entire analysis process.

Modified Critical Decision Method

Many CTA techniques have been developed and used for tasks that involve the practitioner making decisions and taking a course of action based on these decisions, e.g., firefighters, tank platoon leaders, structural engineers, paramedics, and design engineers. A goal

for many CTA techniques is to elicit information on actions taken and the decisions leading up to those actions. However, the IA's job does not fit this pattern of making decisions and taking action/s based on these decisions. One finding that emerged during the initial phase of this research was that making decisions is not a typical part of the IA's task. The major tasks consist of sifting through vast amounts of data to filter, synthesize, and correlate the information to produce a report summarizing what is known about a particular situation or state of affairs. Then, the person for whom the report is produced makes decisions and takes actions based upon the information contained in the report.

A modified version of the critical decision method (CDM) was developed and used for this task domain where the emphasis is on performing analysis (e.g., comparing alternative hypotheses) versus making decisions and taking a course of action. Thus, interview probe questions provided in the literature (Hoffman, et al., in press) were tailored to capture information on IA's approach to gathering and analyzing information. Domain-specific probes were developed to focus the discussion on a *critical analysis assignment* where the analyst had to produce a report on intelligence of a strategic nature. Examples of such strategic analysis problems might include assessments of the capabilities of nations or terrorist groups to obtain or produce weapons of mass destruction, terrorism, strategic surprise, political policy, or military policy. Interview probes were developed to capture information on the types of information used, how this information was obtained, and the strategies used to analyze this information.

CDM. A structured set of domain-specific interview probes was developed specifically for use with the second group of participants. One interviewer conducted the initial interviews; each interview lasted approximately one and one-half hours. Once the initial interview was transcribed and analyzed, the participant was asked to return for a follow-up interview. All three interviewers were present for the follow-up interviews with this second group of intelligence analysts. This approach, requiring two separate interviews, was necessitated by the domain complexity and the desire to become grounded in the case before proceeding with the second interview where our understanding was elucidated and refined.

Deepening Probes. Domain-specific cognitive probes were developed to capture information on the types of information the IA was seeking, the types of questions the

analyst was asking, and how this information was obtained. Additional information was collected on mental models used by analysts, hypotheses formulated and the types of products that are produced. Table 1 lists the questions posed to the participants during the initial interview. Topics for which participants conducted their analyses included modernization of a particular country's military, whether there would be a coup in the Philippines and the potential impact on the Philippines if there was a coup, and the role for the newly created Department of Homeland Security.

Table 2. Modified Critical Decision Method: Deepening Probes

Probe Topic	Probe
Information	What information were you seeking, or what questions were you asking? Why did you need this information? How did you get that information? Were there any difficulties in getting the information you needed from that source? What was the volume of information that you had to deal with? What did you do with this information? Would some other information been helpful?
Mental Models/ Schemas	As you went through the process of analysis and understanding did you build a conceptual model? Did you try to imagine important events over time? Did you try to understand important actors and their relationships? Did you make a spatial picture in your head? Can you draw me an example of what it looks like?
Hypotheses	Did you formulate any hypotheses? Did you consider alternatives to those hypotheses? Did the hypotheses revise your plans for collecting and marshalling more information? If so, how?
Intermediate Products	Did you write any intermediate notes or sketches?

Follow-up Probes. Once the data from the initial interviews was transcribed and analyzed, participants were asked to return for a follow-up interview. The goal during this session was to elaborate our understanding of the IA's task. The analyst was asked

to review the timeline produced during the first interview session and to elaborate on the procedures and cognitive strategies employed. Probes used during the follow-up interview are listed in Table 3.

Table 3. Follow-up Probes Used for Modified Critical Decision Method

Probe Topic	Probes
Goals	What were your specific goals at the time?
Standard Scenarios	Does this case fit a standard or typical scenario? Does it fit a scenario you were trained to deal with?
Analogues	Did this case remind you of any previous case or experience?
Hypotheses and Questions	What hypotheses did you have? What questions were raised by that hypothesis? What alternative hypotheses did you consider? What questions were raised by that alternative hypothesis?
Information Cues for Hypotheses and Questions	As you collected and read information, what things triggered questions or hypotheses that you later followed up?
Information Tools	What sort of tools, such as computer applications, did you use? What information source did you use? What difficulties did you have?

Probes included questions about the participants' goals, whether this analysis was similar to other analysis assignments, use of analogues, and how hypotheses were formed and analyzed. Other probes asked about the types of questions raised during their analysis, methods used, information cues they used to seek and collate information, and the types of tools, e.g., computer software, they used to perform their analysis. During this second interview we went through the same intelligence analysis problem with the goal of obtaining additional details to refine our understanding of the entire analysis process. This included the types of information they used, and how they structured their analysis to answer the strategic question they had been assigned.

Table 4. Cognitive Demands Table: NPS#2

Cognitive Demand	Why Difficult	Cues	Strategies	Potential Errors
Synthe-sizing data	• Lack of technical familiarity with different types of data • Domain expertise is needed to analyze each class of data (HUMINT, SIGINT, ELINT, IMAGERY, etc.)	Difficult to know how to weight different kinds of data	Emphasize type of data analyst has experience with, and disregard other data	• Potential for errors • Tendency to focus on type of data analyst has experience with and to ignore data you do not understand
Synthe-sizing data	• No one database exists that can correlate across systems • No one database can correlate all inputs from many different analysts to form one coherent picture	Systems produce different "results," e.g., mensuration process produces different latitude/ longitude coordi-nates from other systems	Different commands rely on different databases in which they have developed trust	• Users develop comfort level with their system and its associated database; this can lead to wrong conclusion
Synthe-sizing data	• Databases are cumber-some to use: Poor correlation algorithms • System presents results that users do not trust, tracks are "out of whack."	Users don't always understand infor-mation system presents. Too many levels in system are not transparent	Use own experience	• Rely on trend information
Noticing data	• Time critical information is difficult to obtain • Need to assimilate, verify and disseminate in a short time window	Need to decide whether imagery is current enough to proceed with strike How long has it been there?	Need to rely on other sources to verify current	• Refer to other sources to verify

RESULTS

A description of what has been learned during the first phase of this CTA research with intelligence analysts is presented in this section.

STUDY 1

The ACTA method was used with a group that primarily had experience at the tactical level of analysis, thus the discussion was focused on developing a product to

support operations at the tactical level. Using the ACTA method, participants focused on providing descriptions of the cognitively challenging aspects of the task.

Applied Cognitive Task Analysis

The initial set of knowledge representations for the IA's job (produced using the ACTA method) provided the basis for the more detailed CTA. Table 4 presents an example of one of the formats used to codify the knowledge extracted during the CTA using the ACTA method. This Cognitive Demands Table was produced based on analysis of data captured during an interview with one participant. A Cognitive Demands Table provides concrete examples of why the task is difficult, cues and strategies used by practitioners to cope with these demands and potential errors that may result in response to the challenges inherent in the task.

Table 5 presents an example of a Knowledge Audit, which includes examples of the challenging aspects of the task and the strategies employed by experienced analysts to deal with these challenges. A challenging aspect described by several IAs includes the need for the analyst to understand the capabilities and limitations of the systems employed for collection. Understanding the systems' capabilities is important because the systems used to collect data and the tools used to process data can make mistakes due to conflicting databases, complexities of the system that are not transparent to the user and other human-system interaction issues.

Table 5. Knowledge Audit for Intelligence Analyst: NPS#4

EXAMPLE	CUES & STRATEGIES	WHY DIFFICULT?
Collection Ex: Task involves much technical knowledge coupled with experience.		
Start formulating a picture right away	Know what system can do/ limitations Constantly think about nature of the collection system Ask: What do I expect to see here? Constantly checking all data coming in	• Need to understand systems to assess validity of information • All data is not 100% accurate • Collection systems and processors make mistakes: e.g., radar signatures can be similar
Collection Ex: Need to question all data for validity		
Assess validity of information	Correlate signals with what is already Known. Look for incongruent pieces of information.	• Deluged with signals in dense signal environment
Collection Ex: Constant pressure not to miss any little bit		
Huge amount of raw data	Try to extend the area that is monitored to maintain wide area situation awareness	• Analyst has to find the "little jewels" in huge data stream
Collection Ex: Can't miss the radar contact which is the enemy coming out to conduct reconnaissance, or attack the battle group. Want to know 10-12 hours ahead of time when the enemy aircraft was coming.		
Under pressure not to miss anything	Look at *everything* recognizing that probably 90% is going to be of no use.	• Can't afford to let anything slip by without looking at it
Analysis: Focus on what additional information is needed		
Multiple ways to obtain certain kinds of information	Think about what still need to know	• Need some familiarity with different types of sources • Requesting assets to get information may be expensive and conflict with other ongoing things • Potential political ramifications to requesting asset to get something
Analysis: How to present information to customer		
Interpretation can be challenging	Good analyst drives operations Do not just pass all the information without some level of interpretation included.	• Need to ensure customer will take appropriate action as a result of report • Are almost dictating what customer is going to do
Analysis: Pressure to reduce the time to respond		
Analyst brings a lot of knowledge to situation that goes beyond sensor-to-shooter approach	What is the priority of this target vs. others that are out there? Is it the most important thing to do right now? What has occurred in the past week? 2 months? 2 years?	• Things need to be interpreted in *context*

EXAMPLE	CUES &STRATEGIES	WHY DIFFICULT?
Disseminate/ Provide Reports	**Ex: Time-critical spot reports need to go out to people who need it right away**	
Pick out event-by-event pieces	What does customer need to know	• Need to pass time-critical information right away
Disseminate/ Provide Reports	**Ex: See something they don't expect, doesn't fit an established picture**	
Times when event does not fit in with what analyst has been observing recently	Try to develop coherent picture based on other things that have been occurring in past 1-2 hours. What do I think will happen in the next hour? How does the last one event fit in with all the other recent pieces?	• Need to assess how this fits into slightly bigger picture • More likely to discount information if see something you don't expect
Disseminate/ Provide Reports	**Ex: See something outside a pattern of what expected**	
	Always call operator : "We saw X but here is why we don't think it is necessarily the truth." Look for reasons why it might not be correct	• Need to watch your back (not look bad)
Dissemination: Push vs. Pull Technology		
	Simply pushing reports out to people does not always work Pressure on analyst to ensure all high-level decisionmakers have same picture/ information	• High-level decisionmakers want individual, tailored brief: generates differential exchange of information

Another theme that was addressed by many study participants was the constant pressure not to let anything slip by without looking at it. They described this aspect of their task as trying to find the "little jewels in the huge data stream," while knowing that 90% of the stream will not be relevant. An issue germane to analysis, also reported by several analysts, was the tendency to discount information when they see something they don't expect to see, i.e., to look for confirming evidence and to discount disconfirming evidence. An additional pressure experienced by IAs is the need to ensure the customer will take appropriate action as a result of the report (i.e., you are "almost dictating what the customer is going to do.")

Cognitive Challenges

The remainder of this section summarizes what was learned from the ACTA interviews. The IA task is difficult due to the confluence of several factors, including characteristics of the domain and the cognitive demands levied on analysts. The following paragraphs describe the cognitive challenges involved in performing intelligence analysis.

Time Pressure. Decreasing timelines to produce reports for decision-makers is becoming an increasingly stressful requirement for analysts working at all levels, from tactical through strategic levels. An example at the tactical level is provided by a participant who described how the effect of timeline compression coupled with organizational constraints[6] can sometimes "channel thinking" down a specific path.

An example of time pressure at the strategic level is provided by one participant (from study 2) who had six weeks to prepare a report on a matter of strategic importance when he had no prior knowledge of this area and he did not have a degree in political science. The assignment involved the question of whether President Estrada, of the Philippines, would be deposed as President, and if so, would there be a coup? This assignment was to include an analysis of what the impact would be on the Philippines. Six weeks was the total time he had to gather all the necessary information, including the time needed to develop background knowledge of this area. He began by reading travel books and other ethnographic information. This finding is in accord with those of Patterson, Roth, & Woods (2001), i.e., that analysts are increasingly required to perform analysis tasks outside their areas of expertise and to respond under time pressure to critical analysis questions.

[6] This form of organizational constraint, that channels thinking, has been referred to as the "intelligence-to-please" syndrome, a tendency to produce intelligence estimates that support current policy even though information indicates that policy is failing." (Wirtz, 1991, p.8)

Synthesizing Multiple Sources of Information. One aspect of the IA's task that is particularly challenging involves merging different types of information — particularly when the analyst does not have technical familiarity with all these types of information. As an example, two analysts looking at the same image may see different things. Seeing different things in the data can occur because many factors need to be considered when interpreting intelligence data. Each type of data has its own set of associated factors that can impact interpretation. In the case of imagery data, these factors would include the time of day the image was taken, how probable it is to observe a certain thing, and trends within the particular country.

Multiple sources of disparate types of data (e.g., open source, classified, general reference materials, embassy cables, interviews with experts, military records, to name a few) must be combined to make predictions about complex, dynamic events — often in a very short time window. To accomplish the data correlation process, analysts need to be able to combine seemingly unrelated events and see the relevance. The cognitive challenges involved in synthesizing information from these different sources and distilling the relevance can be especially difficult, particularly when different pieces of data have varying degrees of validity and reliability that must be considered. Furthermore, domain expertise is often needed to analyze each type of data.

Human intelligence, electronic intelligence, imagery, open source intelligence, measures and signals intelligence can all include spurious signals or inaccurate information due to the system used or to various factors associated with the different types of data. Analysts described situations where they gave greater weight to the types of information they understood and less weight to less understood types of information. They acknowledged this strategy could lead to incorrect conclusions.

Coping with Uncertainty. Regarding data interpretation, a strong relationship typically exists between the context in which data occurs and the perspective of the observers. This critical relationship between the observer and the data is referred to as context sensitivity (Woods, Patterson, & Roth, 2002). The relationship between context and the perspective of the observer is an essential aspect of the data interpretation process. People typically use context to help them determine what is interesting and informative, and this, in turn, influences how the data are interpreted. Context sensitivity is the framework a person uses to determine which data to attend to and this, in turn, will determine how the data are interpreted. This relationship between context and data interpretation is the crux of the problem for intelligence analysts: When high levels of uncertainty are present regarding the situation, the ability to interpret the data based on context sensitivity is likely to be diminished.

High levels of ambiguity associated with the data to be analyzed produce an uncertain context in which the analyst must interpret and try to make sense of the huge data stream. For instance, data that appear as not important might be extremely important in another situation, e.g., when viewed from a different perspective to consider a competing hypothesis. In general, people are good at being able to focus in on the highly relevant pieces of data based on two factors: properties of the data and the *expectations* (italics added) of the observer. (Woods, et al). However, this critical cognitive ability may be significantly attenuated for professionals in the intelligence community, as they may not always have the correct "expectations" while conducting their search through the data due to the inherent uncertainty associated with the data.

High Cognitive Workload. One of the most daunting aspects of the IA's job is dealing with the high cognitive workload that is produced when a constant stream of

information must be continuously evaluated, particularly when the information often pertains to several different situations. Relevant items must be culled from the continual onslaught of information, then analyzed, synthesized and aggregated. An additional contributor to the high workload is the labor-intensive process employed when an analyst processes data manually — as is often the case — because many tools currently available do not provide the type of support required by analysts. For example, no one single database exists that can correlate across the various types of data that must be assimilated.

IAs often wind up synthesizing all the information in their head, a time-consuming process that requires expertise to perform this accurately, and something that is very difficult for a junior officer to do. Moreover, it is stressful to perform the analysis this way because they worry about missing a critical piece of data and doing it correctly: "Am I missing something?" and "Am I getting the right information out?"

IAs must assess, compare, and resolve conflicting information, while making difficult judgments and remembering the status of several evolving situations. These cognitive tasks are interleaved with other requisite tasks, such as producing various reports or requesting the re-tasking of a collection asset. A request to gather additional information will often involve use of an asset that is in high demand. Re-tasking an asset can be costly and may conflict with other demands for that asset, thus, tradeoffs must be made regarding the potential gain in information when re-tasking the asset to satisfy a new objective. Potential political ramifications of requesting an asset to obtain data to satisfy an objective must also be considered.

Potential for Error. The high cognitive workload imposed on IAs introduces a potential for errors to influence interpretation. For instance, the potential for “cognitive

tunnel vision" to affect the analysis process is introduced by the high cognitive load that analysts often experience. As an example, they may miss a key piece of information when they become overly focused on one particularly challenging aspect of the analysis. Similarly, the analysis process may be skewed when analysts attempt to reduce their cognitive load by focusing on analyzing data they understand and discounting data with which they have less experience. Additionally, discrepancies regarding interpretation may result when decision-makers at different locations (e.g., on different platforms, different services) rely on systems that produce different results. Moreover, the sheer volume of information makes it hard to process all the data, yet no technology is available that is effective in helping the analyst synthesize all the different types of information.

Data Overload. While data overload is a relatively new problem for the intelligence community, it is a major contributor to making the task difficult. It was once the case that intelligence reporting was very scarce, yet with technology advances and electronic connectivity it has become a critical issue today. A former Marine Lieutenant General, describing the situation in the 1991 Persian Gulf conflict commented on the flow of intelligence: "It was like a fire hose coming out, and people were getting information of no interest or value to them, and information that was (of value) didn't get to them." (Trainor, in Bodnar, 2003, p. 55). Data overload in this domain is attributed to two factors. The explosion of accessible electronic data coupled with a Department of Defense emphasis on tracking large numbers of 'hot spots' that place analysts in a position where they are "required to step outside their areas of expertise to respond quickly to targeted questions," (Patterson, et al., 2001, p. 224).

Complex Human Judgments. Difficult human judgments are entailed when (*i*) considering the plausibility of information, (*ii*) deciding what information to trust, and (*iii*) determining how much weight to give to specific pieces of data. Each type of data has to be assessed to determine its validity, reliability, and relevance to the particular event undergoing analysis. Analysts must also resolve discrepancies across systems, databases, and services when correlation algorithms produce conflicting results or results that users do not trust. Evidence must be marshaled to build their case or to build the case for several competing hypotheses and then to select the hypothesis the analyst believes is most likely. Assessing competing hypotheses involves highly complex processes.

Insufficient Tools. The sheer volume of information makes it hard to process all the data, yet the tools currently available are not always effective in helping the analyst assimilate the huge amount of information that needs to be analyzed and synthesized. Many of the systems and databases available to analysts are cumbersome to use due to system design issues. For example, users don't always understand information presented by the system, i.e., when there are discrepancies across system databases (within the ship, within the service, or across services) or the system presents results that users do not trust, e.g., tracks that don't make sense. Tools currently available for use by analysts include poor correlation algorithms and have too many levels within the system that are not transparent to the user.

Organizational Context. Several themes related to organizational context emerged from the interviews. The first involves communication between the analyst and their "customers" (a term used to refer to the person for whom the report or product is produced). When the customer does not clearly articulate his or her need — and

provide the reasons they need a specific item — the analyst has an ill-defined problem. When the analyst does not have an understanding of the situation that merits the intelligence need this will make it more difficult for the analyst to meet the analysis requirement/s. A second organizational context issue is that a goal for analysts is to ensure that all high-level decisionmakers are given the same picture, or information. Yet, high-level decisionmakers will often demand an individual, tailored brief. This generates a differential exchange of information between the analyst and various decisionmakers.

Organizational constraints are placed on analysts to maintain the "status quo," such that new information is filtered through a perspective of being considered as not falling outside of normal operations. There is pressure not to be "the boy who cried wolf." This is in accord with other findings (Vaughan, 1996) who describe organizations that engage in a "routinization of deviance, as they explain away anomalies and in time come to see them as familiar and not particularly threatening." (Klein, et al., this volume). Finally, there is a perception among analysts of feeling unappreciated for their work: Because people often do not understand what is involved there is a perception among IAs that people question "why do we need you?" This credibility issue results in part because different data in different databases produce discrepancies. Intelligence officers feel they loose credibility with operational guys because of these system differences. We now turn the discussion to present results from analysis of data gathered using the modified CDM.

STUDY 2

The modified CDM method was used with a group of analysts who had experience working on analysis problems at the strategic level. When using the CDM,

the emphasis was on having IAs describe tasks where the focus was on analysis of intelligence in order to produce a report to answer a question of strategic interest. The length of time our second group of interviewees had devoted to the assignments that they described ranged from six weeks to three and one-half years (in the latter case, this time was spent intermittently, while serving on a US Navy ship followed by attending graduate school at NPS).

Example 1: Likelihood of a Coup in the Philippines

In this example the person described his task of having to build a brief to answer a political question regarding whether President Estrada would be deposed from the Philippines, whether there would be a coup, and if there was a coup, what the implications would be for the Philippine Islands? What would be the implications for the US? He was asked to complete this analysis task within a time span of six weeks on a topic that was outside his base of expertise (i.e., the geo-political area).

From the initial search of raw reports he produced an initial profile of what was known. Many additional searches and follow-up phone calls were conducted to fill in the gaps in his knowledge and to elaborate on what was learned during the initial set of queries. This step resulted in producing a large number of individual word files on each political person or key player. These included biographies on approximately 125 people, including insurgency leaders, people in various political groups, people with ties to crime, etc. The information in these files was then grouped in various ways to consider several hypotheses. Next he developed a set of questions to use to work backwards to review all the material from several different perspectives to answer a series of questions related to the main question of interest: Will there be a coup? Will it

be peaceful or not? Will it be backed by the military? Will the vote proceed, or will the military step in, prior to the vote? What is the most likely scenario to pan out?

Schemas

A schema is a domain-specific cognitive structure that directs information search, guides attention management, organizes information in memory and directs its retrieval, and becomes more differentiated as a function of experience. Schemas are a way of abstracting the information that has been found so far into a representation. The schema summarizes the external information by abstracting and aggregating information and eliminating irrelevant information. Schemas are structured to efficiently an effectively support the task in which they are embedded.

Figure 2 depicts the schema used to represent the dual-problem space of various information sources that the analyst researched to develop a comprehensive understanding of the issue. The analyst began, in week one, by reading general background information to develop knowledge on the history and cultural ethnography of the country and also by examining prior Naval Intelligence on the previous history for political turnover in the Philippines. During week two he began contacting Intelligence Centers and reading U.S. Embassy cables, an important source for this particular topic. Although this step provided valuable information, because this material was from a secondary source it had to be corroborated. Thus the analyst had to decide which of these reports were to be given greater emphasis and in which reports he did not have much confidence.

One way the analyst structured his analysis was to sort people according to whether they were pro-Estrada or anti-Estrada, which figures would be likely to drop allegiance to the constitution, and so on. The analyst structured, and re-structured, all

the information to see how it might support various scenarios associated with the analysis questions. For example, if the US invests money, will the country remain stable? How should the US react? What is the most dangerous potential outcome? Most/ least likely?

The analyst had five hypotheses that he used to organize his material. Previous coup attempts that occurred around the time of past-President Aquino were reviewed to examine how the allegiance of these people who were involved in past coup attempts might develop. Voting records provided another way to sort people. For a portion of his analysis he used nodal analysis software to examine relationships between people. He used a whiteboard to play "20 questions" to come up with new questions to pursue. Relationship diagrams were constructed for each scenario and tables were developed to facilitate comparison of hypotheses. Many other sources were examined, such as political figures' ties to certain newspapers to determine which camp they would fall into

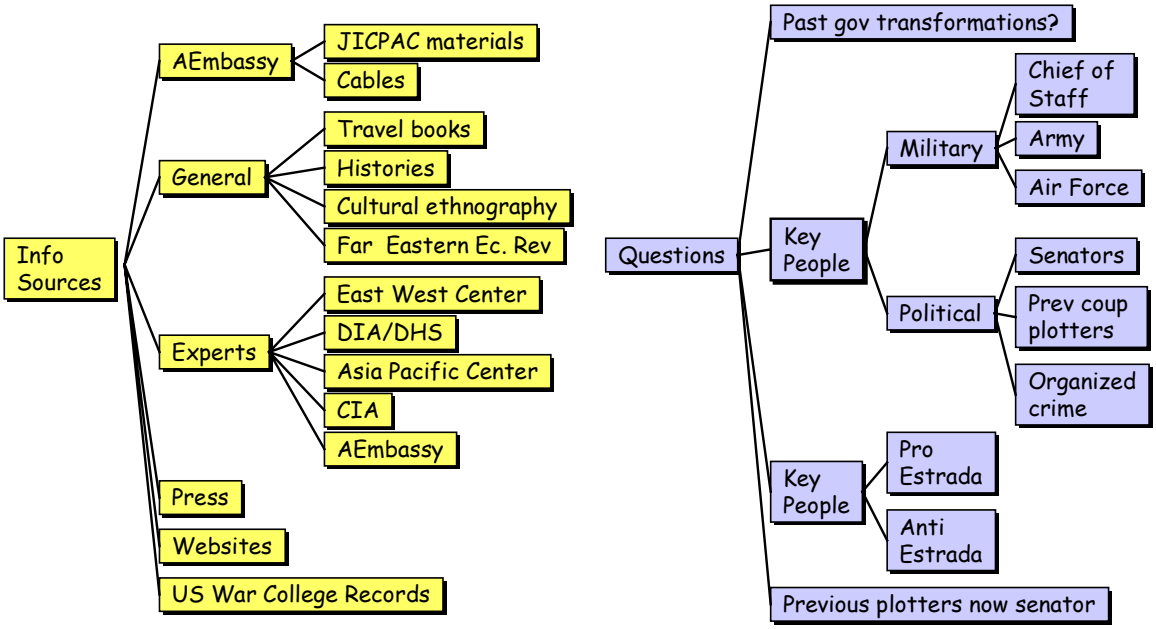

Figure 2. Information Foraging: Dual Problem Space

Figure 3 depicts the information schema used by this analyst. Multiple ways of grouping people were used by the analyst to consider competing hypotheses on how their allegiance would "fall out" based on their various associations. This analyst grouped key people in both the military and civilian sectors according to their military associations, political, family, geographic region, and various other associations, e.g., professional groups and boards they belonged to, to try to ascertain their loyalty. The analyst developed many branches and sequels between people and events in his attempt to examine their affiliations from many different vantage points.

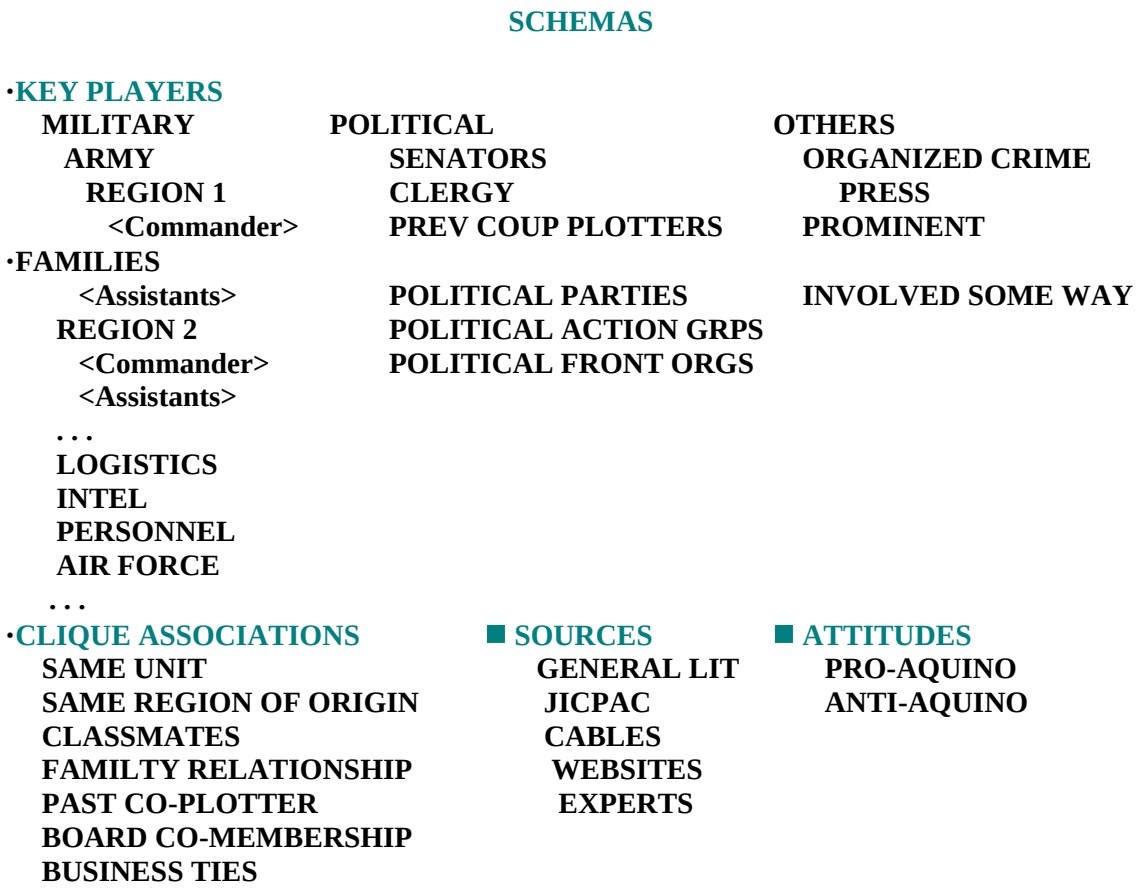

Figure 3. Schemas Used to Analyze the Intelligence Problem

Example 2: Modernization of Country X's Military

This analysis problem evolved as a result of a discrepancy the analyst observed between the stated political military objectives of country X and the observations made by this analyst during a six-month deployment on an aircraft carrier. During his time as strike-plot officer he spent a lot of time collecting and sifting through raw message traffic and interpreting its meaning for the Battle Group. He had developed a considerable knowledge base for this part of the world and was aboard the carrier during the EP-3 crisis, in 2001, when it landed on Hainan Island. During the EP-3 crisis, he was able to provide background information on what had been occurring up to that point as well as during the crisis.

When this analyst reported to NPS to focus on Asia he noticed a disconnect between what professors described in terms of this country's political stance and things he had observed, while operating in this part of the world. Things discussed in his courses were incongruent with the types of military training exercises he had observed this country engage in and the types of military equipment acquisitions made by this country. He began with two or three factors that he knew could be used to support a separate hypothesis to explain the incongruity between what the political leaders are saying and what they are doing. His task was to compare the publicly stated policy of country X regarding their planned military modernization with other possible scenarios for how things might evolve.

This analysis was based on a comparison of this country's officially stated military policy with data collected during detailed observations, and the associated daily reporting, that occurred over a six-month period while the analyst was onboard the aircraft carrier. Table 6 presents a Cognitive Demands analysis of this IA problem. For

analysis of this intelligence problem, the Cognitive Demands analysis described in the ACTA methodology was modified to represent the process that was used by this analyst. Since intelligence analysis involves an iterative process of data analysis and additional collection, we arranged the table to focus on specific data inputs and outputs. Additional columns include cues that generate processes that operate on data, and the strategies or methods used by the analyst to achieve goals when working with specific inputs and outputs. In addition, the table includes expert assessments of why specific inputs and outputs might be difficult. This provides indications of potential leverage points for system design. Finally, the table records specific examples mentioned by the analyst. These examples might be used as task scenarios to guide design and evaluation of new analyst strategies.

Analysis for this task included building the case for several other possible military scenarios regarding actions that might be taken by this country in the future. A comprehensive analysis of two competing hypotheses was developed to take into account future changes in political leadership, the economy, and sociopolitical factors. Data obtained on factors including economic stability, system acquisitions, and military training exercises conducted were manually coded on a daily basis, placed in a database, and aggregated over larger periods of time to depict trends.

Table 6. Cognitive Demands Table for Case 2:
Develop Competing Hypothesis Regarding Military Modernization Efforts of Country X

Inputs	Outputs	Cues/Goals	Strategy	Why Difficult?	Examples
Observations that support hypothesis that country X has embarked on a different modernization effort for a number of years.	Data files that depict country X's trends	Compare stated modernization policy and economic trends within the country	Evaluate the political land-scape of country X, by examining economic and cultural shifts in leadership to gain insight into ways they are looking to modernize.	Stated (public) policy says one thing: Observations point to potentially very different goals.	Types of military training exercises, equipment acquisitions.
Observed modernization efforts	Determine country X's military capability to conduct precision strike	Does the political/ economic/ cultural environment support this operation?	To consider other possibilities beyond their stated military modernization goals	To build a case for possibilities	Use observa-tions from exercises, purchases, etc. to see a different perspective, supported with data
Many prior products: Intel- ligence sources, e.g., unclassified writings, interviews with political leaders	Documents describing discrepancies between observed activities and stated policy.	Notice discrepancies between stated policy and observed activity	Match up things seen in open press with what is occurring militarily	How do observations relate to each other and to the stated policy?	Stated policy of country X does not align with activities observed during exercises.
Classified sources; personal observations; anecdotal memories of deployments and experi-ences from past deployment	Data files of detailed observations gathered over a 6-month period	Help operational side of Navy explore a different view that is not based on established norms of thought	Avoid "group think." Despite the mountain of evidence to the contrary, you don't want to "spool people up."	Difficult to distill the relevance of the informa-tion: Take 100 reports and find the five gems.	Tendency is to report every-thing and treat everything as of equal importance
Read message traffic all day	Two seemingly unrelated events are reported on individually	Take analysis to next level of what is occurring	Ask: "Does this make sense?"	Answer question: "Is this relevant?"	Goes against organizational constraints, i.e., events are "not to be considered outside normal routine training activity."
Volume of information is constrained to the geographic area	Graphs to depict trends of different types of activity	Factor in Army or ground troop movement in addition to Navy activity	Classify infor- mation as relevant or irrelevant. Maintain data-bases of activity, e.g., by day/ week/ months	Several hours a day sorting through message traffic; If had a crisis would be completely saturated.	Group all different categories of activity, e.g., local activity, aggressive activity, exercise activity
Read every-thing can find	1. Brief for the Commander each day 2. Daily Intel Analysis Report	Pick out things that are relevant	Take raw message traffic (w/o anyone's opinion associ-ated with it)	Databases do not match up (even capabilities listed in them)	Extract what think is relevant and highlight activity thought to be relevant
Based on observations of activities that did not match up with what others believed	Form a model of the situation; imagine events over time	To force people to look at a different possibility	Build "Perry Mason" clinch argument	Organization-al constraints not to "go against the grain"	Had lots of documented real world observations
Data on emerging political environment in transition	Understanding of relation-ships between important actors	Paramount to understand who is driving what action	New leadership person is still "driving" things: Added credibility to thesis that there is a split	Could not get access to all material (databases) needed for analysis	Inconsistent capabilities listed in different databases
Location of US forces; geo-political landscape; economic decline affecting country X	Build timeline to depict more aggressive posture	West will not have same influence on economy which leads to political unrest: Political rivalry between old/ new leadership	Describe political factors that could set off a change in direction. Set stage for how things could go in a fictional scenario	Credibility issue: operational guys rarely understand analysis, especially strategic	When presented brief on threat, operational personnel did not perceive information as representative of a threat.
Difference between what they're saying and what they're doing	Revised hypothesis	Initially 2-3 factors that will support a separate hypoth- esis from the accepted hypoth-esis on what is transpiring.	Marshall evidence to support alternate hypothesis	Selecting which pieces of information to focus on	Fact that found so many pieces to support hypothesis indicates hypothesis has to be considered

For this intelligence problem the analyst was looking for evidence to build the case to support several competing hypotheses regarding future political-military scenarios. Several types of information were viewed as indicative of the type of data that could be used to develop and substantiate alternative hypotheses and several methods were used to represent his analysis of the data. For example, a timeline was developed that depicted the following information: (1) location of U.S. forces; (2) geo-political landscape of the world; and (3) the economy, based on economic decline affecting industry in the country. One scenario depicted a situation where the West would not have the same influence on the economy and the fallout will be some political unrest. Political rivalry between the old and new leadership will ensue and the scale will tip to the negative side as a result of political factors that have "gone south." Congressional papers were used, in addition to all the information developed by this analysis, to write a point paper on an assessment of this country's military activity and the kind of threat he saw as a result of his analysis.

Sensemaking

Sensemaking describes one of the cognitive processes performed by the IA to understand complex, dynamic, evolving situations that are "rich with various meanings." Klein, et al, (this volume) describe sensemaking as the process of fitting data into a frame (an explanatory structure, e.g., a story, which accounts for the data) and fitting a frame around the data. The story, or frame, adopted by the IA will affect what data are attended to and how these data items are interpreted. When the IA notices data that do not fit the current frame the sensemaking cycle of continuously moving towards better

explanations is activated. Sensemaking incorporates consideration of criteria typically used by IAs: plausibility, pragmatics, coherence, and reasonableness (*ibid*).

Sensemaking applies to a wide variety of situations. As Klein, et al, describe it, sensemaking begins when someone experiences a surprise or perceives an inadequacy in the existing frame. Sensemaking is used to perform a variety of functions, all related to the IA's job, including problem detection, problem identification, anticipatory thinking, forming explanations, seeing relationships, and projecting the future (*ibid*).

DISCUSSION

Intelligence analysis is an intellectual problem of enormous difficulty (Wirtz, 1991).

Many opportunities for tool development to assist the processes used by IAs exist. Prototype tool development has begun and will continue in conjunction with the next phase of the CTA. Because the ultimate goal is to develop a computational model of the IA's tasks, detailed data must be captured on analysts performing their tasks. Use of process tracing methods, e.g., verbal protocol analysis, in conjunction with the Glass Box software, developed for the NIMD Program (2002), should provide a rich source of data to develop a detailed model of the IA's processes. NIMD's Glass Box is an instrumented environment that collects data on analyst taskings, source material, analytic end products, and analytic actions leading to the end products (Greitzer, 2004)

Use of an instrumented data collection environment in conjunction with think aloud protocol analysis will enable us to gather detailed knowledge about the knowledge and cognition entailed in intelligence analysis. The next phase of this CTA will involve asking SMEs to perform an analysis task while thinking aloud. This

technique typically provides detailed data concerning the mental content and processes involved in a specific task.

Identification of an appropriate sample of problems or tasks is essential to ensure sufficient coverage of critical skills and knowledge. The initial set of interviews was conducted to develop a foundation of knowledge regarding the IAs' task domain. During the next phase of this research additional empirical data will be gathered to further refine the CTA model of intelligence analysis.

Our next phase for this research will involve knowledge elicitation by observing skilled practitioners performing an analysis task using open-source literature. Working within a system development process, to support critical system design issues, additional data and empirical evidence will be collected. The CTA process is an iterative process that builds on subsequent design activities. New tools and training will impact the cognitive activities to be performed and enable development of new strategies. One goal for this phase will be to predict the impact the technology will have on cognition for the intelligence analyst.

REFERENCES

Annett, J. (2000). Theoretical and Pragmatic Influences on Task Analysis Methods. In J. M. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive Task Analysis*, (pp. 25-37). Mahwah, NJ: Erlbaum.

Bodnar, J. W. (2003). Warning analysis for the Information Age: Rethinking the Intelligence Process. Joint Military Intelligence College. Washington, DC.

Chipman, S. F., Schraagen, J. M., & Shalin, V. L. (2000). Introduction to Cognitive Task Analysis. In J. M. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive Task Analysis*, (pp. 3-23). Lawrence Erlbaum Associates, Mahwah, NJ: Erlbaum.

Connable, A. B. (2001). Open Source Acquisition and Analysis: Leveraging the Future of Intelligence at the United States Central Command. Unpublished Masters Thesis, Naval Postgraduate School, Monterey, CA. June 2001.

Cooke, N. J. (1994). Varieties of knowledge elicitation techniques. *International Journal of Human-Computer Studies*, *41*, 801-849.

DuBois, D. & Shalin, V. L. (2000). Describing Job Expertise Using Cognitively Oriented Task Analyses (COTA) In J. M. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive Task Analysis*, (pp. 317-340). Lawrence Erlbaum Associates, Mahwah, NJ: Erlbaum.

Flach, J. M. (2000). Discovering Situated Meaning: An Ecological Approach to Task Analysis. In J. M. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive Task Analysis*, (pp. 317-340). Lawrence Erlbaum Associates, Mahwah, NJ: Erlbaum.

Garst, R. (1989). Fundamentals of Intelligence Analysis. In Ronald Garst, (Ed.) *A Handbook of Intelligence Analysis*. 2nd ed. Washington, DC: Defense Intelligence College.

Greitzer, F. L., Cowley, P. J., & Littlefield, R. J. (2004). Monitoring User Activities in the Glass Box Analysis Environment. Paper presented as part of Panel on Designing Support for intelligence analysts. Human Factors and Ergonomics Society 48th Annual Meeting, New Orleans, LA.

Hoffman, R. R. (1987). The problem of extracting the knowledge of experts from the perspective of experimental psychology. *AI Magazine, 8*(2), 52-67.

Hoffman, R. R., Crandall, B., & Shadbolt, N. (1998). A case study in cognitive task analysis methodology: The Critical Decision Method for the elicitation of expert knowledge. *Human Factors, 40*, 254-276.

Hoffman, R. R., Coffey, J. W., Carnot, M. J., and Novak, J. D. (2002). An Empirical Comparison of Methods for Eliciting and Modeling Expert Knowledge. In *Proceedings of the Human Factors and Ergonomics Society 46th Annual Meeting*, Santa Monica, CA: Human Factors and Ergonomics Society.

Hoffman, R. R., Shadbolt, N. R., Burton, A. M., & Klein, G. (1995) Eliciting knowledge from experts: A methodological analysis. *Organizational Behavior and Human Decision Processes*. *Vol. 62*, No. 2, May, 129-158.

Hoffman, R. R., Coffey, J. W., & Ford, K. M. (in press). *The Handbook of Human-Centered Computing*. Pensacola, FL: Institute for Human and Machine Cognition.

Johnson, R. (2004). http://www.cia.gov/csi/studies/vol147no1/article06.html.

Joint Military Intelligence Training Center, *Open Source Intelligence: Professional Handbook*, Open Source Solutions, 1996.

Jones, M. D. (1995). *The Thinker's Toolkit.* New York: Random House, 44-46.

Klein, G. A., (2001). *Sources of Power, How People Make Decisions*. Cambridge, MA: The MIT Press.

Klein, G., Phillips, J. K., Rall, E. L., & Peluso, D. A. (in press). A Data/Frame theory of Sensemaking. *Expertise Out of Context*.

Klein, G. A., Calderwood, R., & MacGregor, D. (1989, May/June). Critical Decision Method for Eliciting Knowledge. *IEEE Transactions on Systems, Man, and Cybernetics*, *Vol. 19*, No. 3, 462-472.

Krizan, L. (1999). *Intelligence Essentials for Everyone*. (Occasional Paper Number Six). Washington, DC: Joint Military Intelligence College.

Mathams, R. H. (1995). The Intelligence Analyst's Notebook. In Douglas H. Dearth and R. Thomas Goodden (Eds.), *Strategic Intelligence: Theory and Application*, (2nd ed., pp. 77-96). Washington, DC: Joint Military Intelligence Training Center.

Militello, L. G., & Hutton, R. J. B. (1998). Applied Cognitive Task Analysis (ACTA): A Practitioners Toolkit for Understanding Task Demands. *Ergonomics, 41*, 164-168.

Office of the Assistant Secretary of Defense for Networks and Information Integration (OASD/NII) Research Program,

Patterson, E. S., Roth, E. M. & Woods, D. D. (2001). Predicting Vulnerabilities in Computer-Supported Inferential Analysis Under Data Overload. *Cognition, Technology & Work*, *3*, 224-237.

Pirolli, P. & Card, S. K. (1998). Information foraging models of browsers for very large document spaces. In *Advanced Visual Interfaces (AVI) Workshop*, AVI '98. Aquila, Italy, ACM Press.

Pirolli, P., & Card, S. K. (1999). Information foraging. *Psychological Review*, *106*, p. 643-675.

Pirolli, P., Fu, W.-t., Reeder, R., & Card, S. K. (2002). A User-Tracing Architecture for Modeling Interaction with the World Wide Web. In *Advanced Visual Interfaces*. AVI 2002. 2002. Trento, Italy: ACM Press.

Potter, S. S., Roth, E. M., Woods, D. D., & Elm, W. C. (2000). Bootstrapping Multiple Converging Cognitive Task Analysis Techniques for System Design. In J. M. Schraagen, S. F. Chipman, & V. L. Shalin (Eds.), *Cognitive Task Analysis*, (pp. 317-340). Lawrence Erlbaum Associates, Mahwah, NJ: Erlbaum.

The 9/11 Commission Report, Final Report of the National Commission on Terrorist Attacks Upon the United States. 9/11 Commission.

Vaughan, D. (1996). *The Challenger launch decision: Risky technology, culture, and deviance at NASA*. Chicago, IL: University of Chicago Press.

Wirtz, James W. (1991). *The Tet Offensive*, *Intelligence Failure in War*. Cornell University Press. Ithaca, New York.

Woods, D. D., Patterson, E. S., & Roth, E. M. (2002) Can We Ever Escape from Data Overload? A Cognitive Systems Diagnosis. *Cognition, Technology, and Work*, 4, 22-36.

Section 2
Evaluation of a Computer Support Tool for Analysis of Competing Hypotheses

Peter Pirolli and Lance Good

with
Julie Heiser, Jeff Shrager, and Susan Hutchins

Palo Alto Research Center, Inc.

INTRODUCTION

The purpose of this experiment was to evaluate of a computer tool that aims to improve intelligence analysis. This tool provides an external workspace for performing the Analysis of Competing Hypotheses (ACH) method (Heuer, 1999). This experiment focused on a comparison of performance using the ACH computer tool to performance using the ACH method without the computer tool. This experiment was conducted in parallel with a study at NIST (National Institute of Standards and Technology) that focused on an evaluation of the ACH computer tool compared against analysis that does not use the ACH method.

Problems with Intuitive Analysis

Heuer (1999) reviewed psychological literature relevant to the performance of intelligence analysis and identifies various cognitive and perceptual limits that impede attainment of best practice. Human working memory has inherent capacity limits and transient storage properties that limit the amount of information that can be simultaneously heeded. Human perception is biased towards interpretation of information into existing schemas and existing expectations. Reasoning is subject to a variety of well-documented heuristics and biases (Tversky & Kahneman, 1974) that deviate from normative rationality. In problem structuring and decision analysis, people typically fail to generate hypotheses, fail to consider the diagnosticity of evidence, and fail to focus on disconfirmation of hypotheses. ACH is designed to ameliorate the problems with intuitive intelligence analysis that arise from human psychology.

The Method of Analysis of Competing Hypotheses

ACH consists of the following steps.

1. Identify possible hypotheses
2. Make a list of significant evidence for/against
3. Prepare a Hypothesis X Evidence matrix
4. Refine matrix. Delete evidence and arguments that have no diagnosticity
5. Draw tentative conclusions about relative likelihoods. Try to disprove hypotheses
6. Analyze sensitivity to critical evidential items
7. Report conclusions.
8. Identify milestones for future observations

ACH requires that you start with a full set of alternative possibilities (hypotheses) rather than a single most likely alternative. For each item of evidence, it requires you to evaluate whether this evidence is consistent or inconsistent with each hypothesis. Only the inconsistent evidence is counted when calculating a score for each hypothesis. The most probable hypothesis is the one with the least

evidence against it, not the one with the most evidence for it. This is because ACH seeks to refute or eliminate hypotheses, whereas conventional intuitive analysis generally seeks to confirm a favored hypothesis.

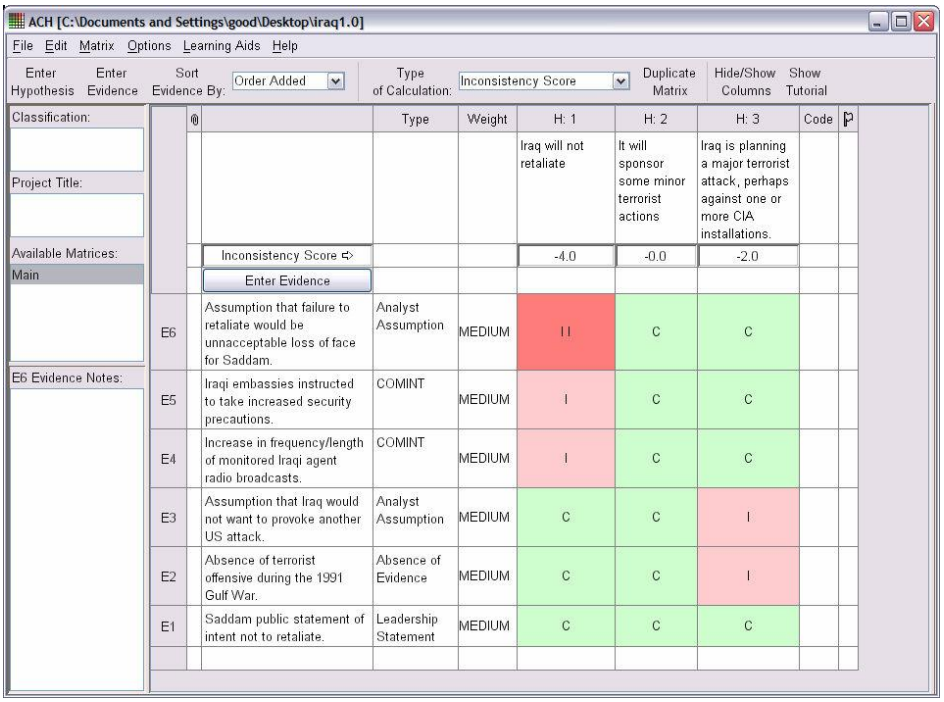

Figure 1. The ACH0 tool containing an analysis example concerning Iraq.

ACH_0

ACH_0 is an experimental program that provides a table oriented workspace for performing the ACH method. ACH_0 allows the analyst to sort and compare the evidence in various analytically-useful ways. It sorts the evidence by diagnosticity, weight, type of source, and date/time. Evidence can be partitioned to compare the probabilities of the hypotheses based only on older evidence versus more recent evidence, or based on open sources versus clandestine sources, or based on the analyst's assumptions and logical deductions versus hard evidence

Figure 1 presents a screen shot of ACH_0, illustrating its table format. The hypotheses under consideration in the example are the columns labeled H1, H2, and H3. Six items of evidence are present in the example in the rows labeled E1 through E6. In the ACH Method, each piece of evidence is assumed to be independent and the hypotheses are exhaustive and mutually exclusive.

An entry of “I” signals that this evidence is inconsistent with the corresponding hypothesis, and entry of “II” signals that it is very inconsistent with the evidence. The “C” and “CC” entries indicate two levels of consistency. ACH_0 distinguishes between “I” and “II” in lieu of a detailed representation of how evidence conflicts with a hypothesis. In other words, it models evidence as being contradictory without saying *how* it is contradictory. (A more detailed representation that focuses on causes of

contradiction could be useful in generating trees of alternative hypotheses). Rather than employing a symbolic representation of contradiction or a probabilistic one, the ACH method simply provides two levels of inconsistency. Similarly, ACH0 provides three levels of weight assigned to evidence. Roughly, this weight is a stand-in for a richer representation of the quality of evidence. Is it reliable? Is the source authoritative? Or is this "evidence" really just an assumption?

ACH0 is intended as a simple tool for organizing thinking about analysis. Its simplicity creates both strengths and weaknesses. Here are some strengths:

- Encourages systematic analysis of multiple competing hypotheses.
- Creates an explicit record of the use of hypotheses and evidence that can be shared, critiqued, and experimented with by others.
- Easy to learn.
- Uses information that analysts can practically understand and enter into the tool.
- Focuses attention on disconfirming evidence – counteracting the common bias of
- focusing on confirming evidence.
- Does not require precise estimates of probabilities.
- Does not require complex explicit representations of compound hypotheses, time, space, assumptions, or processes.
- Works without a complex computer infrastructure

Here are some weaknesses.

- Does not *and cannot* provide detailed and accurate probabilities.
- Does not provide a basis for marshalling evidence by time, location, or cause.
- Does not provide a basis for accounting for *assumptions*.
- Many of the cognitive steps in analysis are not covered at all.

With these caveats ACH0 can have value when used with a clear understanding of its limitations.

Basic Claims about ACH

Three are three key aspects of ACH that are aimed at ameliorating the problems of intuitive analysis above:

1. ACH promotes the generation of a fuller set of alternative hypotheses that each receive equal attention.
2. ACH promotes the identification of key evidence with greatest diagnostic value.
3. Analysts are shaped to seek evidence to refute hypotheses (disconfirm rather than confirm).

These aspects have ancillary effects, including:

- Increasing the odds of getting the right answer
- Providing an audit trail of how evidence used in analysis,

- External matrix representation provides a focus of discussion for a group and can raise disagreements
- External matrix representation increases the amount of information that will receive attention
- Each element of evidence is tested against a broader set of hypotheses
- Greater likelihoods assigned to alternative hypotheses (because of increased attention)

Related Research

A study conducted by Cheikes, Brown, Lehner, and Adelman (2004) investigated the effect of the ACH method in eliminating *confirmation bias* and the *anchoring heuristic*. Confirmation bias (Wason, 1960) is the tendency of people to generate, select, or remember information that confirms a previously held hypothesis. The anchoring heuristic (Tversky & Kahneman, 1974) occurs in judgments under uncertainty when people begin with an estimate of uncertainty (e.g., a probability or confidence rating) and adjust it minimally in light of new evidence. Cheikes et al. (2004) contrasted groups of subjects working with or without ACH on an intelligence problem (Jones, 1995) concerning hypothesized causes of the explosion on the battleship USS Iowa in 1989. Subjects were given three hypotheses to evaluate and received 60 items of evidence. The evidence was delivered to subjects in batches of 15. ACH subjects filled out the ACH matrix with ratings (-2 to 2) of their degree of support of evidence for hypothesis in an ACH matrix. After each batch of 15 items both groups were asked to provide confidence ratings for each hypothesis.

Cheikes et al. (2004) found evidence of a confirmation bias, and also found that non-ACH subjects, more so than ACH subjects, tended to distort their evaluations of evidence to confirm the hypotheses they had been given. There was also a tendency to produce higher confirmation ratings for evidence related to the hypothesis preferred by subjects. This effect was mitigated by ACH for subjects with less analysis experience. Overall ACH appeared to mitigate confirmation bias for analysts with less analytic expertise. While there was some evidence of an anchoring effect, there was no evidence that ACH reduced the effect.

The study presented here was conducted in coordination with one (Scholtz, 2004) conducted at the National Institute of Standards and Technology (NIST). The NIST study evaluated ACH0 with six Naval Reservists. Each subject received two problems and solved the first problem using their normal methodology (control condition) and solved the second problem using ACH_0 (the order of presentation of the specific problems was counterbalanced across subjects).

Overall, the NIST study (Scholtz, 2004) suggests that the Naval Reservists found the ACH0 tool to be useful. Results suggest that

- Analysts were fairly confident of their ability to analyze the scenario.
- Analysts were fairly confident about using the tool but questioned the scoring, and in some cases, the outcome.

- The analysts were able to use the tool with very little difficulty.
- The analysts felt confident that the ACH tool would improve their final report. They were less sure that it would increase the speed at which they could complete the report but felt the tradeoff in quality was worth a reduced speed. Two thirds of the subjects felt that the tool could help show missing evidence to some degree. They were more confident in the ability of the tool to help with the consideration of more hypotheses. They also felt that the tool would help improve the thoroughness of analysis.
- Analysts felt that it would be more difficult to use the ACH method without the tool. They were inclined to use the ACH method in future work and were quite positive about the use of the ACH method in helping them do their jobs.
- Analysts felt the ACH method was easy to learn and use. However, they felt that they had more to learn to apply ACH in their work environment.
- Quantitative data supported the analysts' view that more hypotheses were explored with the tool.
- The analysts did not use the more advanced features of ACH. Only one analyst sorted evidence; few deleted evidence.

METHOD

Participants

Participants were students in an operational intelligence class at the Naval Postgraduate School (NPS) in Monterey. A total of *N* = 25 students participated.

Procedure

Participants were divided into two groups. *N* = 12 students were assigned to the ACH-Computer group and *N* = 13 students were assigned to the ACH-Paper group. Participants in the ACH-Computer group worked individually on cases using the ACH software tool, whereas participants in the ACH-Paper group worked the same cases using paper, rulers, writing utensils, and similar office supplies provided by the experimenters. An attempt was made to match the composition of the two groups by prior experience in intelligence work.

The study took place in a classroom computer laboratory at the NPS campus. The ACH-Computer group worked at personal computers running the Windows 2000 operating system that were arranged facing outwards around the perimeter of the lab. The ACH-Paper group worked at desks arranged in the center of the lab. Instructions were given using PowerPoint slides presented from the from of the lab by one of the investigators. Participants in both groups heard the instructions at the same time, although portions of the instructions might be noted as relevant to only one group or the other. In addition to the investigator presenting the instructions, a second investigator presented an ACH

demo, and three more provided additional assistance. Questions were allowed. Simple questions concerning misunderstanding of the procedure were handled individually. Overall, however, there were less than 10 questions throughout the entire procedure from all students. Almost all of these were simple questions of clarification, although there was one substantive question (see results).

Prior to the experiment, participants were asked to read Chapter 8 of Heuer (1999) which introduces and discusses the ACH method. Participants were also asked to fill out a demographic questionnaire prior to the study. The experiment was divided into two two-hour sessions. Session 1 took place 10:00am – Noon and Session 2 from 3:00pm to 5:00pm.

Session 1 began began with a brief introduction to the experiment and investigators, paper work, handing out of materials, and assignment to groups. Participants were asked to read a brief tutorial. The tutorial for both groups reviewed the ACH method and discussed the analysis of evidence diagnosticity, evidence weighting, and identifying the most likely hypothesis. The ACH-Computer tutorial contained additional information regarding the ACH software tool. Participants were allowed to read through the tutorial until everyone indicated they were done. Following the tutorial, the investigator noted several differences between Chappter 8 of Heuer and the tutorial: (1) hypotheses were to be written out in full, rather than labeled using H1, H2, and so on, (2) a range of consistency labels (CC = very consistent, C = consistent, N = neutral, I = inconsistent, II = very inconsistent, and NA = not applicable) were to be used instead of +/-, (3) a type column was to be used to label the type or source of intelligence, (4) a weight column was to be used with labels HIGH, MEDIUM, and LOW. All participants were given hardcopies of Chapter 8 of Heuer (1999) and asked to turn to the main example used in that chapter, which is an ACH matrix concerning possible reactions of Saddam Hussein during the buildup to the 1991 Gulf War (Iraq Example)Participants in the ACH-Paper group were provided with a practice ACH matrix (with evidence and hypotheses from the Iraq Example already entered) to fill in using the new consistent labels and evidence types. As the ACH-Paper group worked on the Iraq Example, the participants in the ACH-Computer group were asked to open in ACH a file that corresponded to the Iraq example and were provided with a brief demonstration of the menus and features of ACH.

The remainder of Session 1 was devoted to solving Case 1 (Ramos Case). Participants worked on Case 2 2 (JCITA Case) in Session 2. Pilot studies conducted at NIST suggested that Case 1 required 1 hr to complete in full, whereas Case 2 required 2 hr to complete in full. To create a situation with moderate deadline pressures, participants in both groups were given 45 min to construct an ACH matrix for Case 1 and 75 min to construct an ACH matrix for Case 2. Booklets containing descriptions of the cases were handed out, face-down, to all participants, and a “go” signal was given by the investigator to start everyone at the same time. All participants were told when there were 10 min, 5 min, and 1 min remaining. Participants were also instructed to write down their finish time (from a clock on the wall) if they felt they were done prior to the deadline. After the deadline for completing the ACH matrix, all participants were provided with an ACH Reporting form. ACH

Reporting Form required participants to provide a list of their hypotheses for the case, and for each hypothesis to provide a qualitative degree of belief in the hypothesis and a brief (1 – 3 sentence) rationale. Participants were asked to indicate their qualitative degree of belief using one of the following labels: Almost Certain, Very Probably, Probable, Chances about Even, Unlikely, Very Unlikely, and Almost Certainly Not. Participants were given 10 min to complete the ACH Reporting Form.

Following Case 2, at the end of Session 2, participants were asked to complete two forms. The first was the NASA-TLX mental workload instrument. The NASA-TLX instrument obtains measures on six factors involved in overall workload:

- Mental demand, whether the analysis task affects the user's attention or focus
- Physical demand, whether the analysis task affects the user's health, makes user tired, etc.
- Temporal demand, whether the analysis task takes a lot of time that a user cannot afford
- Performance, whether the analysis task is heavy or light in terms of workload
- Frustration, whether the analysis task makes a user unhappy or frustrated
- Effort, whether the user has spent a lot of effort on the analysis task.

Each of the six components is rated on a 7-point scale (1 = low; 7 = high) and each pair of components is compared for difficulty.

The second instrument was a post-test designed to evaluate the ACH tool. This was designed in collaboration with NIST and the details are discussed in the results section.

RESULTS

Problem Structuring

Table 1 presents the mean completion time, number of columns of hypotheses and number of rows of evidence in subjects' final ACH matrices. The experiment was designed to have moderate deadline pressure, so we expected most subjects to use all the time available (Case 1 max time was 45 min; Case 2 max time was 75 min). However, subjects who finished early recorded their completion times. Consequently, the mean times in Table 1 are based on subject data that have a ceiling of 45 min for Case 1 and 75 min for Case 2.

There were no significant differences between the two groups on any of the variables in Table 1: Case 1 Time, $t(15.3) = 1.69$, $p = 0.11$; Case 1 Evidence, $t(19.4) = 1.47$, $p = 0.16$; Case 1 Hypotheses, $t(14.4) = 0.36$, $p = 0.72$; Case 2 Time, $t(16.6) = 0.55$, $p = 0.59$; Case 2 Evidence $t(17.6) = 0.69$, $p = 0.50$; Case 2 Hypotheses, $t(17.8) = 1.36$, $p = 0.19$.

In addition, we examined the proportion of inconsistent ("I" or "II") relations in subjects final ACH matrices. The ACH-Paper group had a mean 25.2% of their ACH matrices filled with inconsistent relations and the ACH-Computer group had a mean 22.6% inconsistent relations, and this was not a significant difference, $t(21) = 0.59$, $p = 0.56$ when computed on the proportions following an arcsine transformation.

In summary, the ACH-Paper and ACH-Computer groups did not exhibit any significant differences on major problem structuring factors that are the main focus of the ACH method: amount of evidence considered, number of hypotheses generated, and amount of inconsistent relations considered.

Table 1

Mean completion times, number of hypotheses generated, and amount of evidence considered (standard deviations in parentheses).

	Case 1 (Ramos)			Case 2 (JCITA)		
Group	Time[1]	Evidence	Hypotheses	Time[2]	Evidence	Hypotheses
ACH-Paper	39.4 (7.2)	11.7 (4.0)	3.6 (1.0)	71.0 (6.9)	21.1 (6.3)	3.8 (1.0)
ACH-Computer	43.0 (2.5)	14.2 (4.1)	3.4 (1.6)	69.2 (7.4)	19.4 (4.4)	4.3 (0.7)

Note: [1]Maximum time in Case 1 is 45 min

[2]Maximum time in Case 2 is 75 min

Post-test Evaluations

Workload

Responses to the NASA-TLX instrument were used to compute overall workload ratings (1 = low; 7 = high). The overall workload ratings were higher for the ACH-Paper group (M = 5.1, SD = 0.91)) than the ACH-Computer group (M = 3.9, SD = 1.34), $t(12.5) = 2.25$, $p < .05$ (computed on log-transformed data). There were no significant differences on any of the six subfactors (mental, physical, temporal, performance, frustration, effort).

Tool Evaluation

Table 2 presents results from the ACH Post-test. The only significant difference between the ACH-Paper and ACH-Computer groups was on the question “How confident were you in your ability to use the tool to perform this task: (Question 5m Table 2, $t(9.4) = 2.33$, $p < .05$ (computed on log-transformed data), which reflects the novelty of the computer tool

Table 2
Mean ratings on ACH Post-test Questionnaire.

Question	ACH-Paper	ACH-Computer
1. Did the task you performed resemble tasks you could imagine performing at work? (1 = not realistic-…5 = realistic)	3.6 (1.5)	3.9 (1.3)
2. How did the task compare in difficulty to tasks that you normally perform at work? (1 = less diffiult…5 = more difficult)	3.3 (1.1)	3.0 (1.4)
3. How confident were you of your ability to analyze the scenario? (1 = less confident…5 = more confident)	3.9 (0.3)	3.7 (1.1)
4. How confident were you of your ability *to use ACH* to accomplish the assigned task? (1 = less confident…5 = more confident)	3.8 (1.0)	3.6 (1.4)
5. How confident were you in your ability to use the tool to perform this task? (1 = less confident…5 = more confident)	**4.7 (0.9)**	**3.2 (1.5)**
6. How would you assess the length of time that you were given to perform this task? (1 = too little…5 = too much)	2.8 (1.2)	3.2 (0.8)
7. If you had to perform a task like the one described in the scenario at work, do you think that using the ACH method would a. Improve your final report? (1=not at all…5=a lot)	4.4 (1.2)	4.0 (0.7)
b. Increase the speed of your analysis? (1 = not at all…5 = a lot)	2.9 (1.1)	3.6 (0.7)
c. Determine which pieces of evidence are critical/missing/inconsistent? (1 = not at all…5 = a lot)	4.4 (0.8)	3.6 (1.2)
d. Enable you to consider and evaluate more hypotheses about the scenario? (1 = not at all…5 = a lot)	3.5 (1.4)	3.8 (1.2)
e. Increase the thoroughness of your analysis? (1 = not at all…5 = a lot)	4.0 (0.8)	4.0 (0.5)
8. Imagine performing this task using ACH but ***without*** a tool to help you organize evidence with your hypotheses. Do you think that task would be (1= easier…5 = more challenging)	4.1 (0.9)	4.3 (0.7)
9. Imagine performing this without using the ACH method. Would this task be easier or more challenging? (1 = not at all…5 = a lot)	3.6 (0.7)	3.8 (0.8)
10. Could you see yourself using the ACH method in your future work? (1 = not at all…5 = a lot)	3.9 (0.6)	3.8 (0.7)
11. Do you think using the ACH method could help intelligence analysts perform their jobs? (1 = not at all…5 = a lot)	4.1 (0.6)	4.1 (0.6)
12. Did you think the ACH method was easy to learn and use? (1 = not at all…5 = a lot)	4.4 (0.7)	4.2 (0.7)
13. After the training you had today, how familiar do you feel with the Analysis of Competing Hypotheses method? (1 = not at all…5 = a lot)	4.1 (0.9)	3.6 (0.9)

General Discussion

Overall, there were few differences between students using the ACH computer tool and students using ACH without computer support. Problem structuring indicators (number of hypotheses, amount of evidence, inconsistent relations) did not differ between the two groups of students. The ACH computer tool did provide improvement in workload.

Student reviews of the NPS course in which the ACH study was conducted were summarized for us by the instructor, who reported:

> I've received the Student observation input from my course this summer. They are anonymous and electronic and are the final view from the students on what worked and didn't from their perspective. I'm pleased to tell you all comments on the ACH evolution were all very positive. Many pointed out the ACH study as a clear highlight for them. In particular, there were several comments about how beneficial the case study was as it forced them to apply in a practical sense some of the theory of what they had been reading and discussing.

These qualitative reviews are consistent with the evaluations provided by the Naval Reservists in the NIST (Scholtz, 2004) study.

REFERENCES

Cheikes, B. A., Brown, M. J., Lehner, P. E., & Alderman, L. (2004). *Confirmation bias in complex analyses* (Technical Report No. MTR 04B0000017). Bedford, MA: MITRE.

Heuer, R. J. (1999). *Psychology of Intelligence Analysis*. Washington, D.C.: Center for the Study of Intelligence.

Jones, M. D. (1995). *The thinker's toolkit*. New York: Random House.

Scholtz, J. (2004). *Analysis of Competing Hypotheses Evaluation (PARC)* (Unpublished Report). Gaithersburg, MD: National Institute of Standards and Technology.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.

Section 3
Collaborative Intelligence Analysis with CACHE and its Effects on Information Gathering and Cognitive Bias

Dorrit Billman
Gregorio Convertino
Jeff Shrager
J.P. Massar
Peter Pirolli

Palo Alto Research Center, Inc.

INTRODUCTION

A recent cognitive task analysis of intelligence analysis (Pirolli et al., 2004) suggests that the analytic process is organized into two interacting major loops of activities: (1) an *information foraging loop* that involves processes aimed at seeking information, searching and filtering it, and reading and extracting evidentiary interpretations and (2) a *sense making loop* (Russell et al., 1993) that involves the development of hypotheses and representations that fit the evidence. This fits a prescriptive view of intelligence analysis that is similar to the idealized practice of science (Kent, 1949), and the problems facing the typical intelligence analyst have much in common with those facing the typical scientist. Typically, the main problem for the process of information foraging is that there are far more data than can be attended to, given the limitations of time and resources. Typically, the main problem for the process of sense making is insuring that a full space of alternative hypotheses has been generated and the alternatives have been tested against the data in an unbiased and methodical fashion (e.g., by adopting a falsificationist methodology and attempting to disconfirm hypotheses). Each of these kinds of problems could be addressed by cooperation or collaboration. Greater amounts of the available information can be foraged if one increases the number of analysts looking at it, assuming that each looks at slightly different subsets of the data. More hypotheses and greater corrective criticism of reasoning would come from having more analysts exchanging their reasoning, assuming some diversity of backgrounds, biases, viewpoints, etc. Indeed, anthropological studies (Sandstrom, 2001) of scientific fields suggest that communities self-organize to form cooperative "informal colleges" of scholars each of whom looks at slightly different phenomena from a slightly different perspective. Unfortunately, one of the recent criticisms of the intelligence analyst community is that it lacks the technology (and culture) required to support such cooperation and collaboration (National Commission on Terrorist Attacks Upon the United States, 2004). The purpose of this paper is present an evaluation of a system called CACHE[7] designed to support collaborative intelligence analysis. The experiment reported in this paper focuses on evaluations of how teams of analysts using CACHE compare to individual analysts using CACHE in terms of information foraging effectiveness and reasoning biases. The experiment focuses specifically on effects on confirmation bias and anchoring because they are considered such significant problems in the evaluation of evidence in the intelligence community (Heuer, 1999).

[7] CACHE stands for Collaborative Analysis of Competing Hypotheses Environment.

Background

Since the early 1970s a great deal of research in social and cognitive psychology has demonstrated that human judgment in decision-making deviates from what is considered normative rationality. Individuals exhibit systematic cognitive biases (Tversky & Kahneman, 1974). Confirmation bias is particularly problematic in intelligence analysis (Heuer, 1999). *Confirmation Bias* is the tendency a decision maker to look for (and give more weighting to) confirmatory evidence, dismiss (and weight less) disconfirming evidence, and use neutral or ambiguous evidence as confirmatory (Kahneman, Slovic, and Tversky, 1982). Related to confirmation bias is the *anchoring effect*, which is the insufficient adjustment of the confidence in an initial hypothesis after receiving new evidence that is inconsistent with this initial hypothesis (Tolcott et al., 1989; Cheickes et al., 2004).

Heuer (1999) proposed a simple methodology called the Analysis of Competing Hypotheses (ACH) to aid individual intelligence analysis, and a simple computer tool (Figure 1) has been developed to support this method. ACH consists of the following steps.

9. Identify possible hypotheses
10. Make a list of significant evidence for/against
11. Prepare a Hypothesis X Evidence matrix
12. Refine matrix. Delete evidence and arguments that have no diagnosticity
13. Draw tentative conclusions about relative likelihoods. Try to disprove hypotheses
14. Analyze sensitivity to critical evidential items
15. Report conclusions.
16. Identify milestones for future observations

ACH requires that one develop a full set of alternative possibilities (hypotheses). For each item of evidence, it requires an analyst to evaluate whether the evidence is consistent or inconsistent with each hypothesis. Only the inconsistent evidence is counted when calculating a score for each hypothesis. The most probable hypothesis is the one with the least evidence against it, not the one with the most evidence for it. This is because ACH seeks to refute or eliminate hypotheses, whereas conventional intuitive analysis generally seeks to confirm a favored hypothesis. The intent of the ACH tool is to mitigate confirmation bias and insure that attention is distributed more evenly across all hypotheses and evidence.

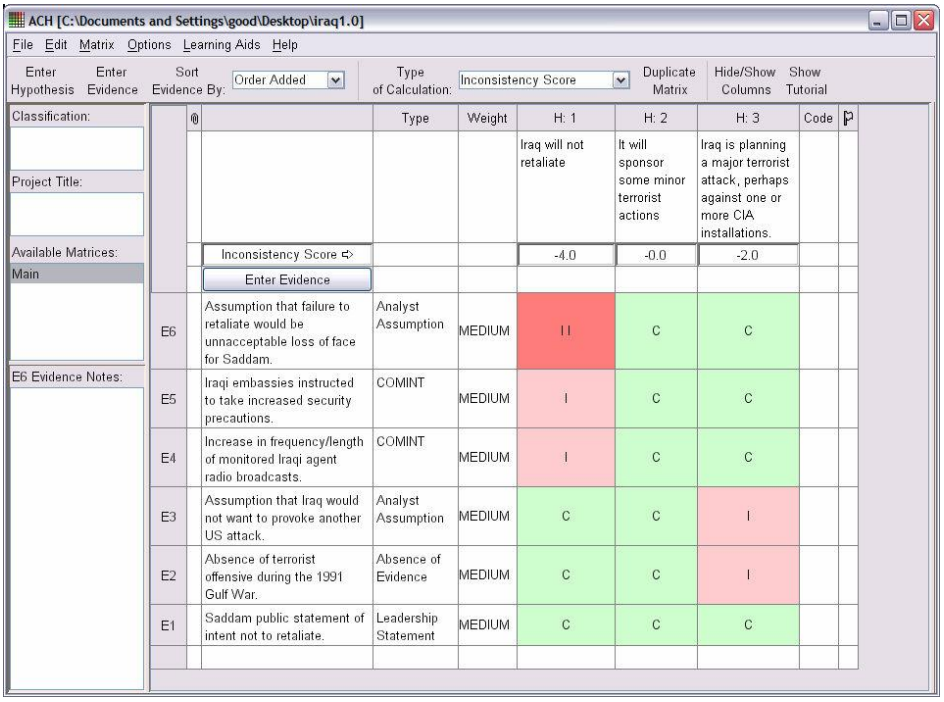

Figure 1. The ACH0 tool containing an analysis example concerning Iraq. Hypotheses are listed in the column headings, evidence is along the rows. The entries in the cells of the table indicate consistent and inconsistent relations between evidence and hypotheses.

The formation of groups to perform decision making is commonly considered a means to accomplishing more thorough processes and less biased outcomes. Consequently, “almost every time there is a genuinely important decision to be made in an organization, a group is assigned to make it – or at least to counsel and advise the individual who must make it” (Hackman and Kaplan, 1974; cited in Nunamaker et al., 1991). In line with the commonly held belief that it is useful to have a variety of views represented in a group, there is substantial experimental evidence that role diversity or functional diversity among group members mitigates bias (Shulz-Hardt et al., 2000) and improves performance (Cummings, 2004). However, there is a large literature suggesting that decision making by face-to-face groups often exhibits the same or worse biases than individual decision making. For instance, there is a tendency for homogeneous face-to-face groups to exhibit confirmation bias in their information search more so than individuals (Shulz-Hardt et al., 2000). This is a specific instance of a more general phenomenon affecting face-to-face groups: their members tend to focus attention on items they have in common and often fail to pool information that has been uniquely attended to by specific individuals (Stasser and Titus, 1985). Research on information pooling in groups has identified several interventions at the level of task and medium that can reduce this group bias (see Stasser and Titus, 2003 for a review). These interventions include the introduction of expert roles (Stewart and Stasser, 1995), the availability of written records (Parks and Cowlin, 1996), framing the decision task as a problem to be solved rather than a matter of judgment (Stasser and Stewart,

1992), requiring members to rank order decision alternatives rather than choosing the best option (Hollingshead, 1996), and more importantly, introducing the technological aid of group support systems (McGrath and Hollingshead, 1994; Dennis, 1996). Several computer support systems for groups have been found to be effective in improving the communication and brainstorming functions of groups, addressing specific process losses that affect face-to-face groups (DeSanctis and Gallupe, 1987; Nunamaker et al., 1991; Dennis and Gallupe, 1993; see Fjermestad, 2004, for a review). For instance, electronic brainstorming systems appear to mitigate biased information search because all information is recorded in an external bulletin board and because factors governing civil face-to-face interaction are lessened.

In general, prior research has suggested that the performance of a decision-making group is affected by properties of the task (e.g., intellectual vs. judgment task; structured vs. unstructured), medium (i.e., face-to-face vs. collaborative system) and group (e.g., diversity of pre-group individual biases and group size). The medium can impact the amount of group bias (Benbasat and Lim, 2000) and the group outcome, especially when a correct answer is missing and response is based on choosing a preferred alternative and reaching consensus (i.e., judgment tasks) (Straus and McGrath, 1994). The composition of a group can affect its performance and this effect can interact with the medium. Heterogeneous group composition introduces in the group the potential for de-biasing itself (Shulz-Hardt et al., 2000). Diversity increases the potential of the group to expose members to different sources of information, know-how, and feedback (Cummings, 2004). However, the costs for translating such group potential into actual group performance might depend on the support provided by the medium of interaction. We should also consider that the presence of different perspective (e.g., roles) and information across the members increases the cost for grounding communication and coordinating (Clark, 1996). However, the role that computer tools can play to improve the quality of judgment in group decision-making remains largely unexplored (Benbasat and Lim, 2001; Lim and Benbasat, 1997).

Study Goals

The high-level goal of this experiment is to investigate effects of computer mediated collaborative intelligence analysis on amount of evidentiary information considered and the mitigation of cognitive bias. In the present study, we manipulate the composition of decision making teams in terms of the diversity of biases held by individuals when they began an intelligence analysis task. Through experimental manipulation, members of a decision-making team could be biased toward the same solution to an analysis problem (*homogenous* group) or to different solutions (*heterogeneous* group), prior to doing the analysis task. We then measure the effects of our manipulations in terms of amount of cognitive bias observed in the analysis outcome and the amount of evidentiary information attended (*information coverage*). Analysis in these groups is compared to individual analysts. We are

interested in evaluating whether (1) heterogeneous groups of analysts produce less cognitive bias than homogeneous groups or individual analysts and (2) if CACHE, our collaborative medium, is able to temper the process cost of group interaction when comparing the performance of interacting groups with the performance of sets of non-interacting individual.

Our high level goal for this study involves three subgoals:

1. *Investigation of bias in groups*: it compares performance across differently composed groups: a Heterogeneous Group, in which group members initially are each biased toward a different hypothesis, a Homogenous Group, in which group members are initially each biased toward a common hypothesis, and a Solo, or Nominal Group, Condition in which the (initially biased) participants work alone. Our experimental design varies the group composition, to assess how task performance supported by CACHE changes across group structure. This goal motivates much of our experimental design, method, and analyses.
2. *CACHE evaluation*: it evaluates CACHE, providing information about how participants use CACHE and what CACHE's strong and weak points were. All participants in the experiment used CACHE, allowing us to collect as much, broad information as possible about its usability. Our investigation in this regard is exploratory, not hypothesis testing.
3. *Experimental method*: it develops and tries out a laboratory method for assessing group processes in judgment tasks. An important feature of our procedure is that the judgment tasks are completed by each individual group member and not by the whole group. This enables comparisons between conditions with interacting and nominal groups, at the cost of reducing partially the collaborative nature of the task.

As part of this research, we calibrated material, task procedure, and evaluation metrics and we informally report on the method's successes and weaknesses.

CACHE

CACHE is the platform used to support collaboration in this study. CACHE builds upon the ACH method discussed above. A variation of the earlier ACH0 interface (Figure 1) is at the heart of CACHE (Figures 2 & 3). CACHE employs a a simple decision matrix, where the rows represent evidence and the columns hypotheses. Each cell of the matrix represents the relationship between one piece of evidence and one hypothesis. In the particular implementation described here, the cells may take on values of CC (very consistent), C (consistent), N (neutral), I (inconsistent) or II (very inconsistent). Internally these take on numerical values of: CC=2, C=1, N/A = 0, I = -1, and II= -2. The use of these values will become clear momentarily. In addition to cell values relating evidence to hypotheses, each evidence item (each row in the matrix) may be given a "weight" from among: Very-Low (1/6), Low (2/6 low), Medium (3/6), High (4/6), and Very High (5/6).

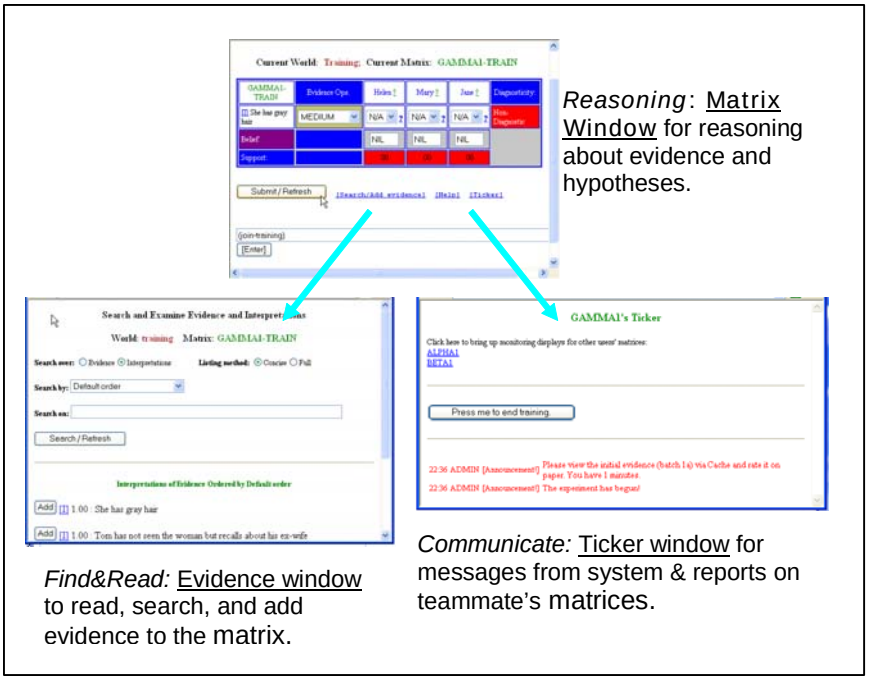

Figure 2. Three of the CACHE: the user's primary decision matrix, the evidence window showing interpretations of each piece of evidence, and the ticker window which reports system events and reports every time a partner posts evidence to their matrix not in the user's own matrix.

CACHE is a client-server system which is used through a standard web browser (i.e., Internet Explorer). A given subject (whom we shall always refer to as user "Alpha" here) works in collaboration with two other subjects ("Beta" and "Gamma" – Alpha's "co-workers"). Alpha may have open some of the following sorts of windows:

1. A *primary decision matrix* in which selected evidence is recorded, and in which the relation between each piece of evidence and each hypothesis is indicated (as above);
2. One or two windows that enable Alpha to view (but not change) the matrices of Alpha's co-workers (that is, views of Beta and Gamma's matrices);
3. A *search interface* over the evidence that is available to Alpha and which works much like Google's search interface (this evidence may be different for Alpha, Beta, and Gamma);
4. A *ticker* that, when the experiment is in "collaborate" mode, indicates when Alpha's coworkers select pieces of evidence that Alpha has not used;
5. A number of *evidence viewers*, each examining the details of one piece of evidence;
6. A *chat window* enabling broadcast communications among all subjects and with the experimenter.

Figure 2 shows schematically a screen with three linked windows. For collaborating users, many windows might be open at once, particularly if a user kept partners' matrices open. Users were instructed to always keep the main matrix, chat, and ticker windows open.

On the server side a state machine drives CACHE through the phases of the experiment. In each phase three subjects in an analysis team can see certain types of evidence, and have certain available operations. For example, in the middle phases, each subject in a homogeneous group has access to similar evidence, whereas the three subjects in a heterogeneous group see have access to different evidence.. Similarly, at the beginning of each phase the subjects must work alone. (I.e., they cannot open viewers into one another's matrices, and the ticker does not report the actions of one's coworkers.) In the latter part of each phase, the ticker is activated, and users can view one another's matrices. (In fact, they are told to do so.)

Each matrix in the experiment had the same three hypotheses (referred to here as "Friction", "Overraming", "Suicide", described in detail below). Only the evidence for or against these hypotheses can be changed. When a user choose to add a piece of evidence to a matrix, s/he could then (and usually should) set the weight of that evidence, and the values of the cells between that piece of evidence and each of the three hypotheses to indicate the level of support that the new piece of evidence gives to each hypothesis. At the bottom of the matrix (see Figures 2&3) the user can also set his or her *subjective* certainty in each of the three hypotheses, represented in terms of integer percentages which sum to 100. Each time settings of this sort are completed, two important actions take place automatically: First, the matrix interface calculates the objective level of support for each hypothesis. This is simply the sum of the cell support values (-2 to +2) times the weight of each piece of evidence (1/6 to 5/6). This, then is normalized to 1.0 over all three hypotheses and the score displayed at the bottom of the matrix (see Figure 2).

The second thing that happens is that if the subjects are in a collaborative phase (where they can see one another's matrices, and where the ticker is running), the ticker reports to each user what his or her coworkers have changed in their matrix. Specifically, it reports when a coworker has added evidence, deleted evidence, and when cell settings of evidence that both subjects share are different (called here "inconsistency").

The ticker offers the ability for the user to examine the evidence noted in the ticker entry. For example (again from Alpha's point of view), if the ticker reports that Beta has added a piece of evidence that Alpha simply does not have access to it currently, Alpha can examine that evidence by clicking the link in the ticker window (or in the view window onto Beta's matrix). Once Alpha has done this through any means available, that evidence is now available to Alpha as well, even if it was not previously available according to the order of operation of the experimental phases. This is one way that collaboration makes new knowledge available: Once a co-worker has used some piece of evidence, a user (e.g., Alpha) is given the opportunity to use it as well, even if the user did not previously have access to it.

Every operation carried out either by the user or by the server (e.g., moving through the phases, some of which are timed) is logged. These logs provide data for .Summary of Hypotheses

In this CACHE study we experimentally manipulated cognitive bias in individual subjects, and manipulated the mixture of biased individuals forming analyst teams (groups). Our measurements focus on the detection of differences in information gathering (information coverage) and detection of changes in cognitive biases. The main focus of our analysis was on the effects of working collaboratively in a group (as opposed to working solo) and the effects of working in groups with different mixtures of individual biases. We hypothesized that

- Heterogeneous groups would show *less judgment bias* than Homogeneous groups. Because CACHE supports sharing information among participants, the differing views in the heterogenous groups should mitigate cognitive biases and facilitate exposure to greater amounts of evidence (greater information coverage) relative to homogenous groups.
- Heterogeneous groups would show *no net process loss* relative to the Solo analysts. We expect that CACHE, the collaborative medium, will mitigate the process costs of group interaction.

EXPERIMENTAL DESIGN

These hypotheses were tested with subjects using CACHE to solve an analysis task that involved reading through material about the real case of the explosion in one of the 16-inch gun turrets on the battleship USS Iowa that occurred in April 1989, and assessing the relative likelihood of three hypothesized causes of the explosion:

- **Hypothesis 1: An overram ignited powder**. An unqualified rammerman inadvertently caused a mechanical rammer to explode powder bags
- **Hypothesis 2: Friction ignited powder** by causing a buildup of static electricity inside the gun chamber causing a spark that ignited the powder
- **Hypothesis 3: Gun captain placed an incendiary device** in order to purposely kill himself and others.

Experimental instructions and materials were designed to systematically induce an initial belief favoring one of these three hypotheses about the cause of the Iowa explosion. This initial bias was induced by assigning roles to individuals. Subsequent manipulation of the presentation of evidence about the case was performed to initially reinforce these biases, but by the end of the experiment participants had access to a collection of evidence that was carefully balanced to equally support each of the hypotheses. Thus, at the end of the experiment the unbiased or the normatively correct judgment would be a balanced distribution of belief among the three explanations: all are similarly supported by evidence (33%, 33%, 33%).

Table 1. Distribution of participants across Initial Belief and Group Condition.

Group Condition / Initial Belief	Nominal Group	Homogeneous Group	Heterogeneous Group	Total
Hypothesis 1 (bias)	1 group (3 S_S)	1 group (3 S_S)	1 group (3 S_S)	3 groups (9 S_S)
Hypothesis 2 (bias)	1 group (3 S_S)	1 group (3 S_S)	1 group (3 S_S)	3 groups (9 S_S)
Hypothesis 3 (bias)	1 group (3 S_S)	1 group (3 S_S)	1 group (3 S_S)	3 groups (9 S_S)
Total	3 groups (9 S_S)	3 groups (9 S_S)	3 groups (9 S_S)	9 groups (27 S_S)

The initial bias toward one of the hypothesis served as a baseline, from which belief change and final belief could be assessed. As shown by prior research on judgment bias, people frequently overweigh prior belief, over-anchor on initial judgments, and focus on confirming rather than disconfirming evidence (Tversky and Kahneman, 1974). Our goal was to determine if our manipulations of collaborative analysis mitigated these biases and improved information coverage.

The independent variables in this experiment were:

a. **Group Condition**. This was a between-subject factor with three levels: Homogeneous Group, Heterogeneous Group, Solo Group. Each group was composed of three individuals, interacting (Heterogeneous and Homogeneous) or working alone (Solo).

b. **Initial Belief**. This was a between-subjects factor, used for counterbalancing. It was orthogonal to Condition: In each condition a third of the participants had each of the three values of Initial Belief. In the Homogeneous Condition, individuals in the same group all had the same Initial Belief, with belief changing between groups. In the Heterogeneous condition, individuals in the same group each had a different initial belief. In the Solo Group individuals did not interact so variation of Initial Belief may equivalently be thought of as within or between group. For our labeling purposes we grouped three individuals with the same Initial Belief into the same Solo Group. The distribution of participants across Condition and Initial Belief is shown in Table 1.

c. **Block.** This was a within-subject factor, Evidence was presented in four blocks and blocking is described in the Method section.

The dependent variables in this experiment were Bias and Information Coverage. We measured bias (or debiasing process) in several, increasingly fine-grained ways.

1. We directly assessed the **final beliefs**. This provided the simplest, most direct measure of bias. Final beliefs could be compared both across conditions and to a normative distribution of belief. If participants in all conditions were similarly biased initially, then differences in final belief would reflect different debiasing processes among conditions.
2. We measured the **change in belief** (degree of debiasing) from the belief expressed at the end of the initial bias-inducing procedure to the final belief expressed at the end of the entire judgment task. This measured, individual by individual, the degree (and direction) of belief change. We also considered belief at intermediate points. These first two types of measures provided a "bottom line" of bias, from all contributing processes. They incorporated effects of any anchoring bias from the process of forming and "committing to" an estimated value for initial belief, or any confirmatory bias in how later evidence was selectively consulted and incorporated, and of any other biasing processes.
3. We separated the underlying **judgment processes** that contributed to bias. Of particular interest, we assessed *how evidence was used and the distribution of evidence use* between confirming and disconfirming evidence, and across evidence relevant to each of the three hypotheses. We looked at measures reflecting how information was weighted and integrated in reaching judgment and at what information participants read. This third, process-oriented measure provided information about whether and when confirmation bias guided the selection or use of evidence. The information coverage of individuals and groups was calculated on the basis of these results.

These three measures provided information about judgment bias due to over-reliance on initial beliefs and provide ways to assess the anchoring effect and confirmation bias discussed above.

METHOD

Participants

The participants were recruited among graduate and undergraduate students at Stanford and PARC in the summer 2005. Thirty-three students participated in the experiment and were assigned to 3-member same-gender groups (11 groups). 2 of the 11 groups were excluded from the final data analysis because of irregularities in the procedure or technical problems. The final experiment sample comprises 27 participants (9 groups). About 2/3 (17/27) of the participants were males and 1/3 (10/27) were females, the average age was 25.2; only two participants were older than 40 years.

Setting

The three members of each group were seated at workstations located in separate rooms. They could not see each other and were able to talk to the experimenter through a chat tool. The members of interacting groups could also share information with their partners using the chat tool and the collaborative components of CACHE. Participants in the same group were trained together, and thus had slight familiarity with one another.

Apparatus

Each participant completed the task using the CACHE system and the chat tool (Figures 2 & 3). The CACHE system comprises a suite of tools supporting collaborative decision-making ([1]). The tools supporting the analysis of each user are the ACH matrix, search page, and read/interpret page. The ACH matrix provides a table-oriented workspace for performing decisions using a structured method for analysis - the Analysis of Competing Hypotheses (ACH) method. CACHE also includes tools supporting collaborative analysis: the ticker, which ireports on differences in the evidence included by the teammates in their matrices, and two read-only views of the entire partners' matrices. The chat tool, paired with the CACHE system, enabled synchronous communication and coordination about the task.

The participants used the Analysis of Competing Hypotheses (ACH) method for performing their analysis task. This method was developed by the intelligence community for structuring the analysis process and enhancing the quality of decisions about complex decision-making tasks (Heuer, 1999). It helps the decision-maker to assess if (and how strongly) the available evidence supports or refutes the hypotheses that are inherent in arguments. To apply this method, each participant was provided with a set of alternative hypotheses and a large body of evidence. They were asked to evaluate whether and how strongly multiple evidence items were consistent or inconsistent with each hypothesis. Two general rules were emphasized: (1) analyze all the evidence with respect to all the hypotheses, and (2) emphasize disconfirming evidence. These drive the overall process and are critical to determining the relative likelihood of the competing hypotheses (Cheikes et al., 2004).

[1] CACHE includes other collaborative features that were not used for this experiment.

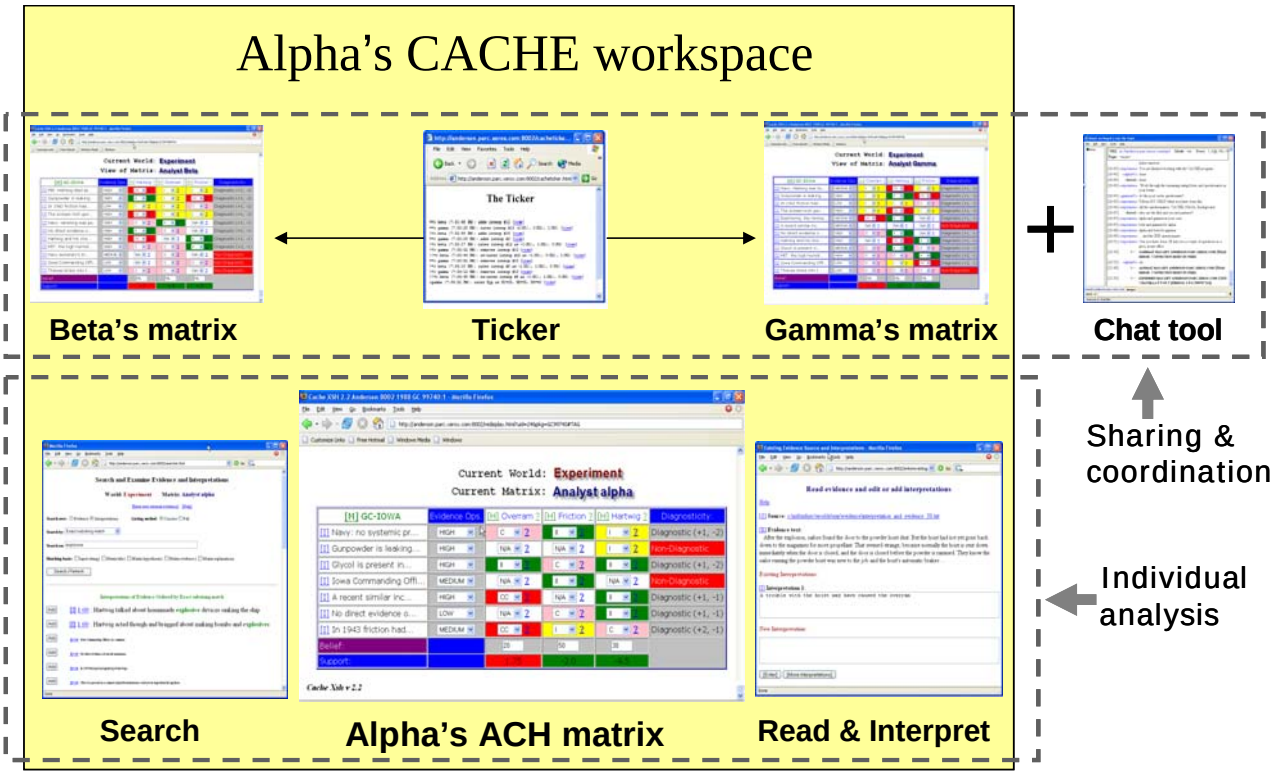

Figure 3. CACHE Workspace for one team member (Alpha): Search, ACH matrix, and Read & Interpret page for individual analysis (bottom); read-only views of partners' matrices, Ticker, and Chat tool for team coordination (top).

Task

The experimental task was to analyze the Iowa Explosion investigation case: assess the relative likelihood of three hypothesized causes of the explosion on the battleship USS Iowa in April 1989. The participants were asked to analyze the hypotheses and four batches of evidence using the ACH method and the CACHE system. The case includes three hypotheses and 80 pieces of evidence, 45 positive items supporting one of the hyptotheses, 15 negative items disconfirming an hypothesis, and 20 neutral fillers. The evidence items come from a variety of sources, e.g., results and opinions provided by government investigators, independent testing organizations, and subject-matter experts engaged as consultants.

The Iowa Explosion case had been used as the principal exercise for practicing the ACH method in Jones (1998, pp. 209-216) and was adopted by Cheikes et al. (2004) to study confirmation bias in individual analysts. In adapting the prior task materials (Cheikes et al., 2004) to the collaborative context three main changes were made: the use of a greater number of evidence items, the use of interpretations in combination with each evidence item, and the use of professional roles that favored specific sources of evidence. Roles were assigned to group members consistently with the other information manipulations in order to induce bias. The amount of positive (supportive) vs. negative

(disconfirming) evidence items and tagged vs. untagged (with the source name) evidence items were balanced across the three alternative hypotheses.

We used 80 pieces of evidence and 80 interpretations. The analysis of evidence items and hypotheses was conducted in 4 blocks. In each block the decision-makers had access to 20 new evidence items and 20 new interpretations. Each evidence item was 1 or 2 paragraphs in length and was summarized by a 1-sentence interpretation. It was assumed that each interpretation had been entered by an analyst, who had read the evidence item before. The interpretations were designed to show whether the corresponding item supported, opposed, or neither supported or opposed one of the hypotheses. The evidence covered various topics (e.g., mechanics, electronics, and psychological diagnoses) and contained conflicting expert testimony connected to different sources of evidence (Navy experts, Sandia labs, FBI experts), as is typical of complex analysis tasks (law enforcement investigations, intelligence analysis, emergency management, and financial analyses).

Overall, the analysis of the Iowa Explosion case is a structured decision-making task (the ACH method) conducted with the support of CACHE (see Nunamaker et al. (1991) for the benefits of using structured tasks in computer-supported groups). The content and type of information available to each decision-maker was controlled by the experimenter. The key parameters that were controlled include number of evidence items and interpretations supporting/opposing each hypothesis, type of evidence item (i.e., positive vs. negative, tagged vs. untagged) and source of evidence (Navy, Sandia, FBI). Aspects of the information that were actively manipulated by the experimenter in order to induce bias were:

- The professional roles of analyst (Navy, Sandia, or FBI expert) favoring one of the sources of evidence
- The ordering of evidence supporting the three hypotheses within a block (block 1);
- Relative proportion of evidence supporting the three hypotheses within a block (block 2);
- Independently rated relevance of evidence supporting the three hypotheses within a block (blocks 1-4);

Procedure

During the first portion of the experiment the participants signed the informed consent form and received training and background information. They were trained, as a group, on the ACH method and the CACHE tool, and were given the opportunity to practice with both the method and tool. The training lasted about 35 minutes. After a short break, they sat at workstations located in separate rooms. They were given the instructions about the task and the role and background information about the Iowa Explosion case. After they had read the background information they started the analysis task.

The task was organized so that the analysis of the case was decomposed into four blocks. During each block each participant received through CACHE a block of interpretations and evidence items. S/he was instructed to search, read and make sense of the evidence; and to add to their ACH matrix in CACHE the evidence items considered relevant. For evidence items entered in the matrix, s/heindicated within the matrix the degree each evidence item supported, or conflicted with each of the hypotheses; indicated the importance of that piece of evidence, and indicated at the bottom of the matrix the overall level of confidence in each hypothesis at the end of each block. CACHE also generated and displayed at the bottom of the matrix a linear strength-of-evidence measure derived from the user's entries.

During the first block all the participants worked alone. Then, for each of the remaining three blocks the members of interacting (Homogeneous and Heterogeneous) groups collaborated remotely with their partners in addition to performing their individual analysis in CACHE. At the beginning of each block they worked individually for five minutes, before interacting with their team members. In contrast, the members of Solo (nominal) groups worked individually for the entire duration of the task. The analysis of the case lasted about 1 hour and half. At the end, the participants were administered a questionnaire and a short interview.

RESULTS

EFFECTS OF GROUP CONDITION ON BIAS

This results section reports on a) the existence of bias, b) effects of condition on overcoming biases from over-reliance on initial beliefs, and c) what we have learned about the judgment process in this task. We have three main measures: 1) Final matrices, 2) Degree of Belief entered by users into their matrices at the end of each block, and 3) the computer logs preserving user interactions with CACHE.

Our design was to collect a very large number of measures. However, we have missing data on several measures, which influences the strength of conclusions we can draw. We have complete data, for each user, on the Final Matrix. For the Degree of Belief recorded at the end of each block, we are missing 1 user's data from Block 1, 2 users from Block 2, and 4 users from Block 3. We have Cache-logs for all three Homogeneous groups, 2 of 3 Heterogeneous groups, and 2 individual solo users.

Many of the analyses concern bias. Call the hypothesis towards which a particular user was initially directed, the Preferred Hypothesis. Several measures can reflect bias, which can in turn be tested for interaction with Condition. The bias measures we use are:

a) Response to the preferred hypothesis, such as degree of belief in the Preferred Hypothesis; call this the direct measure

b) An interaction effect between Initial Bias (between users) and response hypothesis (within user, e.g. judged importance of evidence supporting each of the three hypotheses). Belief in the three response hypotheses should be higher for the hypothesis towards which the user was initially biased, producing the intereaction.
c) A derived difference measure.

To derive the difference measure, subtract the average response to the Alternative Hypotheses from the response to the Preferred Hypothesis,. Values greater than zero indicate bias. Effects of condition can be measured as main effects (e.g. a main effect of condition on degree of belief in the Preferred Hypothesis), as interaction with a difference score, or as a three-way interaction , Condition and Initial Bias (between-subject factors) with the within-subject variable distinquishing the responses to the three hypotheses.

Initial Bias

To assess effects of condition on overcoming bias, we had to be able to produce controlled initial biases in all conditions. We used two dependent variables to assess the existence and nature of bias at the end of Block1. First, we predicted that the degree of belief in an hypothesis would be affected by the bias manipulation, but not condition. Recall that through the end of Block 1, there was no difference in the way participants in different conditions had been treated. The Belief for each hypothesis at the end of Block 1 are shown in Figure 4. S. A repeated measures ANOVA shows a strong interaction between rated hypothesis and bias manipulation, $F(4,34)=8.49$, $p<.001$, but no effects of Condition or any of its interactions $F<1$.

Second, we used the direct measure, here Belief entered in the matrix for the Preferred Hypothesis. If, averaged across counterbalancing of content, more than a third of users' belief is committed to the preferred hypothesis, then users were successfully biased. The overall % Belief in Preferred hypothesis, and the values for each condition (Heterogeneous Condition mean=56, Homogenous Condition=57, Solo=62) all differed from an unbiased 33%. Neither Bias, Condition, nor their interaction had significant effects on Belief Preferred at the end of Block 1. Thus we succeeded in biasing our users, and doing so very similarly across conditions.

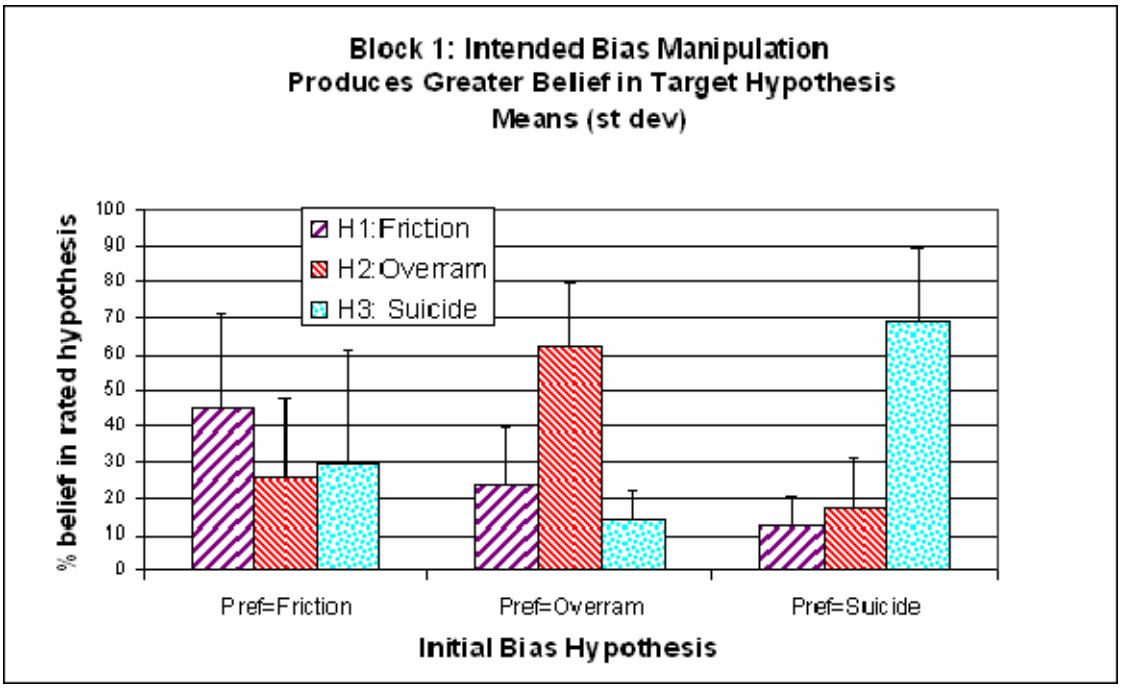

Figure 4. Manipulating bias towards preferred hypothesis does produce greater belief in that hypotheses. No difference among conditions at the end of Block 1.

Effect of Condition on Bias Change

Do users in some conditions reduce bias more than those in other conditions? In particular, has Belief in the Preferred Hypothesis become closer to the normative value of 1/3 in some conditions more than others? Descriptively, Figure 5 shows that the initial % Belief in Preferred Hypothesis changes over time, with the Solo and Heterogeneous Conditions diverging from the Homogeneous Condition. Belief in Preferred Hypothesis increases from 57% to 72% for the Homogeneous Condition (increasing bias), while it decreases in the Heterogeneous and Solo Conditions from roughly 60% to 43%.

We conducted ANOVA's for testing the effect of condition on the Belief in the preferred hypothesis at each of the 4 blocks. Because of our pattern of missing data, this approach has much more sensitivity than repeated measures ANOVA or MANOVA. Effect of Condition was not significant at Block 1 ($F=.13$, $p=.874$) but was significant or marginally significant at Block 2 ($F=3.8$, $p=.041$), Block 3 ($F=2.93$, $p=.076$), and Block 4 ($F=3.92$, $p=.034$). The reduced significance of the effect of Condition at block 3 may be due to the greater data loss for this block. As a complementary analysis we did a MANOVA on Belief at each of the four blocks with condition as a factor. Here Condition overall was not significant, $F(8,32)=1.51$, $p=.193$, nor significant on Block 1 , $F(2.18)=..41$ 13, $p=.670$. It was, however, significant or marginally significant on the remaining blocks (Block 2, $F(2,18)=3.38$, $p=.057$; Block 3, $F(2,18)=3.22$, $p=.064$; and Block 4, $F(2,18)=3.70$, $p=.045$).

Recall that more observations are missing in the intermediate blocks (2 in Block 2 and 4 in Block3), making these points less reliable.

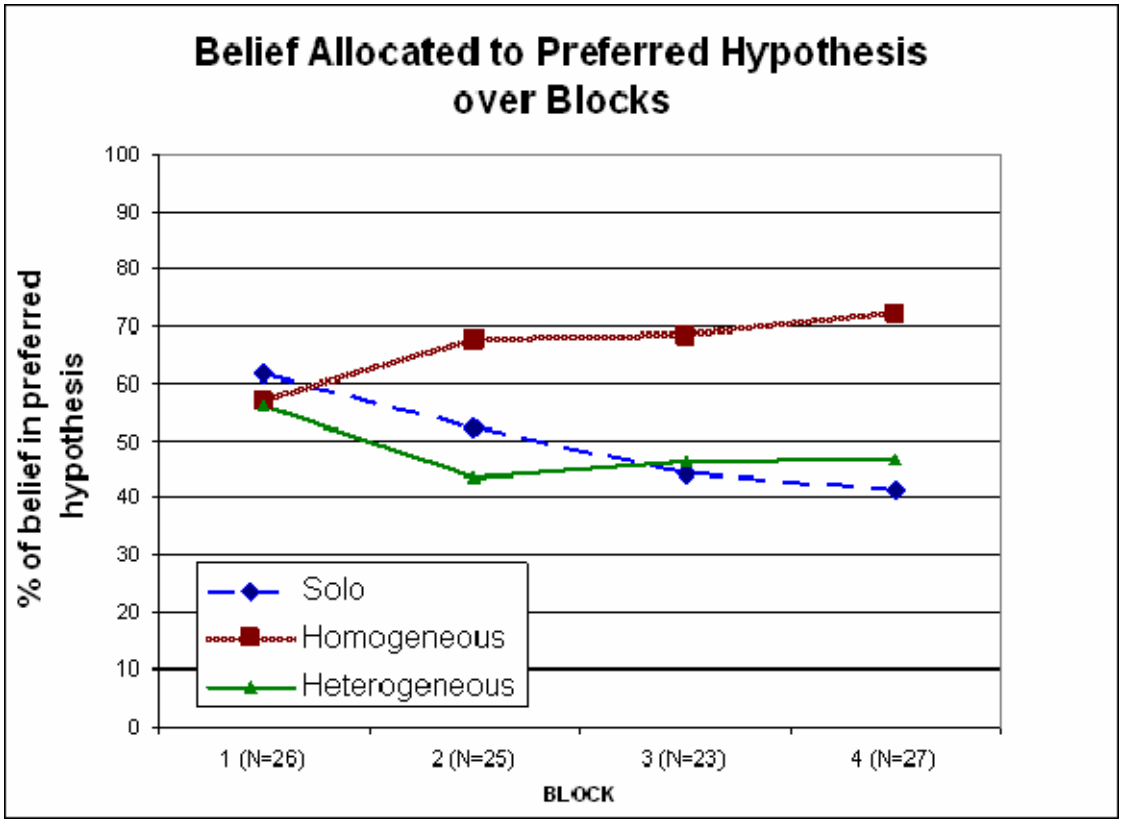

Figure 5. Percent of belief allocated to the preferred hypothesis (matrix hypothesis toward which the individual user was biased) at the end of each block. Standard Deviations for Homogeneous and Heterogeneous Conditions. N = number of users with data at each time point.

Processes in Judgment and Bias

Given the presence of biased judgment and differences among conditions in reducing bias, we can begin to separate constituent influences on judgment. We report results loosely ordered from earlier processes, such as inclusion of an item in the matrix, to later judgment steps, such as final belief.

A. Final Matrix Data

Evidence Counts: Inclusion in the Matrix. Our evidence pool included items with disconfirming information, directly undermining support for one hypothesis. Because negative information, whether or not disconfirming a favored hypothesis, may be difficult for people to use, we checked the rate of inclusion of positive versus negative evidence. Overall, users included 68.5% of the 15 pieces of negative evidence and 69.1% of the 45 pieces of positive evidence; neither Condition, Item Positivity, or their interaction were significant in a repeated measures ANOVA. This strikingly similar proportion of negative to positive evidence used is of interest in its own right. It suggests that instruction in the ACH method with its emphasis on the value of disconfirming information may have been sufficient to overcome biases against use of negative information.

Table 2. Final Matrix Measures with test of Bias and Condition effects

			CONDITION means *(st dev)*			
Type of Response	**Specific Measure**	**Sig.**	**Hetero**	**Homo**	**Solo**	**Tot. Ave.**
Amount Included in Matrix	# of Pieces of Evidence		42.33	33.67	43.78	39.93
			19.47	*11.25*	*13.12*	*15.14*
Amount Included in Matrix	Positive Evidence Difference Score		1.61	2.17	1.17	1.65*
			1.95	*2.63*	*1.70*	*2.09*
Importance	Positive Evidence Difference Score	+	1.07	1.51	1.17	1.25*
			0.21	*0.47*	*0.20*	*0.36*
"Value" (Evidence +/- Rel. to Hypothesis)	Difference Score		3.06	8.22	1.72	4.33*
			7.92	*7.69*	*10.77*	*9.02*
Support (summed Importance X Value)	Difference Score		12.11	38.78	12.89	21.26*
			26.78	*29.41*	*42.08*	*34.51*
Strength (CACHE integrated measure)	Strength of preferred hypothesis	+	3.67	6.75	-1.33	2.88*
			3.80	*3.48*	*8.58*	*6.56*
Belief (User entered judgment)	Belief in preferred hypothesis	+	46.67	72.22	41.44	53.44*
			27.16	*15.64*	*29.75*	*27.61*

+ Significant effect of condition
* Significant bias. Test for bias varies with measure; see text.

Given the similar inclusion rates of negative and positive evidence and similar total amounts of evidence used across conditions, we asked if there was bias in how evidence was selected, and whether there were any differences across conditions in this. Confirmation bias would be expected to show up as greater inclusion of positive evidence supporting the Preferred versus alternative hypotheses. The effect of bias on negative evidence might be more complex. Bias towards a particular hypothesis might either increase attention and inclusion to any evidence relevant to the hypothesis, or discourage inclusion of any contradictory evidence.

We used the interaction term between Initial Bias and Matrix Hypothesis as the indicator of bias. We ran a repeated measure ANOVA with Matrix Hypothesis and Positive/Negative evidence type as within-subject factors and Condition and Initial Bias as between subject factors. The main effects of Positive/Negative, $F(1,18)=186.9$, $p<<.001$ reflected the much greater amount of positive evidence. There was a main effect of Matrix Hypothesis, $F(2,17)=4.813$, $p=.022$, but not of Initial Bias or Condition. The marginally significant interaction of Matrix Hypothesis and Initial Bias, $F(4,36)=2.53$, $p=.057$ showed some evidence of bias. In addition, item type (Positive/Negative) interacted with the bias indicator, in the three-way interaction, Positive/Negative X Matrix Hypothesis X Initial Bias, $F(4,36)=5.79$, $p=001$. This suggests that Positive and Negative items were biased differently and should be assessed separately. Positive items showed a significant bias effect (interaction $F(4,36)=5.35$, $p=.002$), but no condition or Matrix Hypothesis effect (F's <1.6, $p>.2$). More positive items were included for the Matrix Hypothesis which matched the Initial Bias hypothesis. No effect of bias on negative or neutral items were found. In summary, while there is some bias in how positive items are selected for inclusion in the final matrix, this does not differ with condition. Neither amount nor distribution of evidence included is significantly affected by Condition; however, on both, the Solo Condition trends toward more favorable performance.

We also asked whether groups, collectively, covered different amounts of evidence in different conditions. For example, if the three individuals in the heterogeneous condition looked for different types of evidence but combined what they found, they might together use more diverse evidence and produce broader coverage than three independent individuals. For solo, or nominal, groups, we grouped the nine individuals into three groups in two ways, putting users with the same bias together and regrouping to put users with different biases together. Average group coverage by the three Heterogeneous Groups (mean=63.7/80), three Homogeneous Groups (mean=52/ 80), and six nominal/solo groups (three of homogeneous (mean=64.7/80) and 3 of heterogeneous (mean=67/80) members) was similar, and quite high. The data suggests a disadvantage of homogeneous members, but no advantage of interacting vs nominal group.

Evidence Importance. A second point at which confirmation bias might affect judgment is weighting the importance of the evidence which has been included in the matrix. Users' ratings of

evidence importance were converted to a 1-5 importance scale. The summed importance ratings of a particular class of evidence might be influenced by confirmation bias. An overall bias to give more importance to evidence relevant to the Preferred hypothesis would be expressed as an interaction between Matrix Hypothesis and Initial Bias. This interaction was significant, $F(4,36)=3.276$, $p=.022$, in a repeated measures ANOVA with Condition and Initial Bias as between- subject factors and Matrix Hypothesis within subjects. There was no significant main effect or interaction with condition, p's >.3; main effect of Matrix Hypothesis was significant, $F(2,17)=4.159$, $p=.034$.

The effects of a confirmation bias on positive evidence are clearer, and possibly different from the effects on negative evidence. Therefore, we ran the parallel analysis on positive evidence alone. Here we found a strong, significant bias (Matrix Hypothesis X Initial Bias), $F(4,36)=12.48$, $p<.001$ and also a significant interaction of this bias with Condition (3-way interaction, $F(8,36)=2.343$, $p=.039$). To understand this bias, and translate an interaction into a main effect, we also used the difference score: (the summed importance of preferred-hypothesis evidence) – (average of summed importance of the two alternative-hypothesis evidence). Table 2 shows this difference score. An ANOVA on the difference score shows significant condition differences in amount of bias (condition $F(2,18)=6.00$, $p=.010$; and effect of Initial Bias, $F(2,18)=2.42$, $p=.027$). The Heterogeneous Condition is significantly less biased than the Homogeneous Condition in summed importance of evidence (Post hoc, Tukey HSD, $p=.008$)

Evidence Value: Relation of Evidence to Hypotheses. Assigning the relation between evidence items and hypotheses is a third point at which confirmation bias might affect judgment. Users responses from CC to II were coded on a 1-5 scale, where 5 was highly consistent and 3 was neutral.

We analyzed the summed values of evidence in several ways, all evidence together, separately for positive evidence, for neutral evidence, and for the difference score on summed value. Presence of bias would produce significant interaction of Initial Bias and Matrix Hypothesis, and several analyses did have this interaction significant. However, the pattern of interaction was not directly that predicted by a confirmation bias. Nevertheless, the difference score bias measure did differ from zero, $t(1,26)=2.497$, $p=.019$. suggesting some reliable bias. No analysis showed a reliable influence of condition.

Evidence Support: Value * Importance. We multiplied evidence weights times evidence values, to derive an integrated measure of support. (This is very similar to the CACHE generated strength measure for each hypothesis displayed to users at the bottom of their matrix, as shown in Figures 2&3.) To assess overall bias, we again constructed a difference score, subtracting the average support for the alternative hypotheses from the support for the favored hypothesis.

Mean support by condition is shown in Table 2. Though the differential support for the biased hypothesis, mean=39, in the Homogeneous Condition is a larger value than the differential support,

mean=12, in the other conditions, variability is high. We found evidence of bias (1-sample t-test compared value to zero, t=3.20, p=.004). There was no significant effect of condition in the one-way ANOVA on the difference measure, F(2,26)=1.85, NS, nor was there in any of several ANOVA's and MANOVAs, testing support from all evidence and separately for positive and neutral evidence. (Tests on all evidence and positive evidence found the interaction effect indicating bias, as well.)

Beliefs. Analysis of Belief (Table 2), in the final matrix were presented in the results section on bias change, establishing the existence of bias at the end of the experiment and the emergence of condition differences. Further, we compared the CACHE-derived Strength measure and found condition differences in degree of bias on this measure as well (Table 2). For most but not all users, the ordering of degree of belief across the three hypotheses was the same as the ordering derived by linear combination of weight and value of evidence. By and large, the final judgments seem, indeed, to derive from the steps of entering and evaluating evidence which we requested users to do. Degree of correspondence between these two metrics is an interesting question for further work.

B. Cache Log

For 5 groups of users and 2 solo users we had the CACHE-log records of their interaction. This provided additional information about the processes, in particular, the number of times a user accesses the full evidence, not just the interpretation, presumably in order to read it. We tallied these as access-to-read. The 9 Homogeneous condition users read 34.7 items on average, compared to 45.5 items averaged by the 6 Heterogeneous Condition Users. The CACHE-logs also record when users add evidence to their matrix; this can't be less than the number of pieces of evidence in the final matrix, though users can delete evidence after adding it. Homogeneous users averaged 39.2 items added, compared to Heterogeneous user average of 53.2.

The pattern of Heterogeneous users adding (and reading) more evidence than the Homogeneous users is found here as well. These data allow us one new comparison: between amount read and amount added. Users add more evidence than they read, in both Homogeneous and Heterogeneous conditions. This is consistent with a 'breadth first' strategy, of processing a large amount of information superficially, and including unread material in the matrix. We had not anticipated users developing this strategy.

During Blocks 2 and 3, each user has access to some evidence presented only to his or herself. If another user accesses this information, they must have done so through a collaborative window, either a partner's matrix or the ticker (and not the user's own evidence pool). These events are a particularly interesting point of collaboration, and the CACHE-log also stores all such Borrow events. Groups and individuals varied enormously. The two heterogeneous groups had 3 and 53 Borrow events, with Borrowing distributed across users and evidence. The Homogeneous Groups had 2, 5,

and 16 Borrow events, with the Borrowing in the high use group primarily one user to read and reread a small set of evidence.

C. Confidence

Level of confidence in judgment might be a mediating process in debiasing. Shulz-Hardt et al. (2000) observed that the differences in judgment bias between homogeneous and heterogeneous groups could be mediated by the level of confidence within the group. Consensus within a group (higher in homogeneous groups) increases the level of confidence; this heightened confidence tends to reduce the willingness to engage in effortful processing and search for new information

In our study, we asked the participants to indicate how certain (on a scale from 0 to 8) they were about the correctness of their decisions at the end of each block. We tested for differences in the level of confidence across the three group conditions and across blocks. We expected to observe a higher level of confidence in homogenous groups than in heterogeneous groups. This difference was expected to appear after the group members were allowed to collaborate with their partners (after block 2): a significant interaction between Block and Condition. A repeated measures ANOVA with Group Condition and Block (1 and 4) as independent variables shows that Block (1 and 4) had a significant effect on the level of certainty ($F(1, 24)=6.95$, $p=.014$) but the interaction between Block and Condition was not significant. Therefore, the significant difference observed was due to Block and not to Group Condition;this does not confirm the findings of Shulz-Hardt et al (2000).. Participants in all group conditions tended to gain more confidence as they examined a larger quantity of evidence during their analysis of the case.

D. Summary.

We have shown a) that initially, at the end of Block 1, users are biased as intended, b) that initially users in the three grouping conditions are very similarly biased, c) that at the end of the experiment the groups differed in their bias, specifically, the bias of the homogenous groups was high (and increased) while the biases of the homogenous and solo groups were lower (and decreased over exposure to additional evidence). On all measures, whether or not significant differences among conditions were found, this pattern was found: worst performance in homogeneous groups and similarly better performance in the solo and heterogeneous groups.

Although the user could simply write in any values for belief, normatively, there are several prior activities which should and apparently do contribute to the final belief judgment. Because CACHE and the task encourage and record these activities, we can begin to localize the component judgments where bias is most visible and where conditions differ most.

We could reliably detect bias in inclusion of positive evidence in the matrix, in evidence importance, in the combination of evidence importance and value, and in the final belief. We could

reliably detect condition influences on bias in rating the importance of positive items and in the final beliefs. We have not addressed the issue of how much we detect the influence of condition where we do because the influence is most potent for the processes reflected in these measures, and how much we detect where we do because these measures have less noise. Nevertheless, our findings suggest that the primary way in which group structure mediates bias change may be at the point of assessing the importance of pieces of evidence, specifically that evidence relevant enough to be included in the matrix. Weighting the importance of evidence to be greater when it supports a prior belief can be one expression of a confirmation bias. In turn, group structure may influence confirmation bias through influencing the assessment of evidence importance.

Usability Questionnaire Ratings & Free Response

Assessment of Work Context

The questionnaire provides information about the task and supporting elements in which CACHE was used. The most important result here is the time pressure experienced by participants in both group conditions, but not in the solo condition. Evidence comes from both ratings and free response.

Ratings on the three task-level questions, on time available to perform the task, confidence in doing the task, and resemblance to real-world tasks, were assessed in a MANOVA, Condition $F(6,46)=3.261$, $p=.009$. This effect was driven by differences in ratings of too little or too much time (Condition $F(2,26)=8.27$, $p=.002$). Participants in the Solo Condition averaged 3.2 (SD=.83), close to a “3” indicating neither too much nor too little time. This contrasts with the ratings of too little time in the Homogeneous (mean =1.9, SD=.78) and the Heterogeneous (mean = 2.1, SD=.60) Conditions. Participants in interacting, versus nominal, groups experienced more time pressure, presumably a result of the additional process costs of managing group interaction.

Support for the greater time pressure experienced by the interacting groups also comes from the free response measures. Though none of these open questions mentioned the task or experiment, 8 of 18 interacting users commented on time-pressure or inadequate time, while none of the 9 Solo users did so.

A 6-question MANOVA assessing the ACH method itself and the Chat tool found no condition differences ($F<1$), and ratings were generally positive. We read free responses for information about task environment and experiment context. One frequent comment was on insufficient training and practice with CACHE before beginning.

Experienced Usability of CACHE

CACHE was used successfully in this task, as judged by users’ comments and the Usability Questionnaire. In general users were engaged, found the task interesting and rated CACHE

positively. For all 35 questions, a rating of 5 marked the positive end of the scale. We analyzed responses in several ways

We formed six composite measures (using 26 of the 35 questions) on 1) improving performance, 2) learnability of components, 3) ease using different components,4) usefulness of different components, and 5) helpfulness on different subtasks. In addition, we took an average of these averages to get an overall indicator of usability.

The overall average of ratings, for all users, was 4 .04 on the 5-point scale, indicating general acceptance and experienced usability of CACHE. Descriptively, the Solo Condition is slightly more favorable, and the Homogeneous Condition less favorable, on many measures. A one-way ANOVA on the Overall Rating found a marginal effect of condition, $F(2,24)=3.22$, $p=.058$, but a MANOVA on the six composite measures with condition as a factor did not show significance, $F(10,42)=1.55$, $p=.156$. On all six composite measures, the confidence interval for each condition's mean was above 3, the neutral midpoint of the scales.

We selected the 23 questions which focused on CACHE as opposed to the task or ACH method, and which were applicable to solo as well as group conditions. Thus, these ratings on these questions can be seen as repeated assessments of CACHE. A repeated measures ANOVA (condition as factor, item type repeated measure, $F(2,23)=4.56$, $p=.021$) but not the MANOVA (23 dependent variables, Condition as factor, $F(44,6)=2.10$, $p=.177$, found significant difference between conditions.

These descriptive and inferential statistics show a broad acceptance of CACHE, common across conditions.

Condition Differences and Points for improvement

In addition to global assessment we wanted to know a) which aspects of CACHE most differentiated conditions, and b) which aspects of CACHE most merited improvement. We used a consistent standard to identify condition differences at the item level: $p<.05$ on the item's univariate comparison calculated as part of the large, 23-response MANOVA reported above.

One item stood out as receiving ratings which were both very low on average and which differed across conditions, $F(2,23)=5.867$, $p=.009$: usefulness of the ticker window. Participants in both Solo (means of 2.3) and Homogeneous (mean 2.2) conditions rated the ticker window very low, both relative to the Heterogeneous Condition (mean 3.8), and in absolute terms. This was the only item on any aspect of the system which was rated under 3. For participants in the Solo Condition, the ticker only provided information redundant with experimenter-provided chat, about progress through the experiment. Hence the ticker probably served little function here. However, the ticker for Homogeneous as well as Heterogeneous participants posted messages about partner activity, so this use cannot explain the Homogeneous users dislike for the ticker. Further, Homogeneous and Heterogenous users rated "value of information from partners" very similarly (means of 3.7 and 3.9,

respectively). Users in the Homogeneous Condition may feel less urgency about using partner information, and wish not to be interrupted to attend to it.

In addition, conditions differed on rating usefulness of two other components: the window for searching and adding interpretatons, and the window for reading the underlying evidence and adding interpretations. The ease of learning the window for reading the underlying evidence and adding interpretation was also rated differently across conditions. One global assessment differed between conditions: confidence in using CACHE to accomplish the task. On these four measures, Solo users rated favorably and Homogenous users less favorably.

Recommendations about CACHE

Three broad areas emerged as areas for improvement. Interestingly, none of these was directly assessed in the rating questions we had designed. First, many users (14/21, distributed across conditions) commented about the need for better tools to order or manipulate evidence in the matrix. Users wanted to be able to flexibly reorganize evidence, for example, ordering by relevance to one hypothesis, grouping contradictory evidence together, and sorting by topic. Users also found it hard to locate newly added evidence in the matrix. On a slightly different point, some users who had added their own interpretation of evidence found it confusing to track what evidence was entered under what interpretation.

Second, ten users asked for improved window management and a reduced number of windows. This was a particular issue for participants in the Heterogeneous Condition (7 of 9 users), probably because these people were making heaviest use of partner matrices, thus increasing the complexity of their window management task.

Third, users had varied questions and issues about use and coordination of the multiple ratings. It is not clear how much of the difficulties expressed would be addressed with more extensive training, and how much they reflect unwanted costs from manipulating and viewing so many distinct ratings.

There were three user generated ratings, and all provoked comment. Seven users expressed concerns about the ratings of evidence relevance, particularly lack of clarity in the labels used and in the role of the values entered, and also frequently forgetting to enter values because of confusion with the default value (Medium). Our users commented about difficulty using the CC/II ratings of relation between evidence and hypothesis. One user explained that s/he didn't know when to use NotApplicable versus Inconsistent, particularly if a piece of evidence was irrelevant to a hypothesis.

To our surprise, the only user mentioning Belief ratings, our primary dependent variable, questioned their value as "just my gut feeling." Other users may have felt similarly, because we found that it was quite difficult to get users to update the belief values in their matrices when requested to do so.

On the primary, system-generated measure "support," four users said it was very valuable (3 in the Solo Condition) and five commented about not understanding what it meant or how it was

derived. On the system-generated measure of diagnosticity, ten users commented negatively on the comprehensibility or usefulness of the ratings; indeed, due to the secondary importance of this information, we had not allocated any training time to this.

How Does CACHE Support Collaborative Judgment

High Volume of Evidence Processed

CACHE facilitated using large amounts of evidence. Overall, users incorporated 40 of the 80 pieces of evidence in the roughly 70 minutes users had to work with their evidence matrices. This is a large amount of information considered. We did not include a No-CACHE condition, so we cannot make direct claims about CACHE's role supporting this performance. However, a cross-study comparison with Pirolli et al (Nov 2005) is suggestive. They tested a related tool, ACH0, which provides a similar evidence matrix but requires users to type in evidence and hypotheses themselves. The studies differ in multiple aspects: problems were different, the ACH0 users added their own hypotheses as well as evidence, and our users had been introduced to their problem before they began work with the evidence matrices. Nevertheless, comparison is informative. In the ACH0 study, users had 45 minutes and entered an average of 14 pieces of evidence on one problem and took 75 minutes in entering 19 pieces of evidence for a second problem. Indeed, based on this study, we designed our task with the intent of giving our users substantially more evidence than they could be expected to process.

This contrast in amount of information entered in the matrix when users do or do not have to type in the evidence suggests that allowing users to click-to-add evidence produces a dramatic jump in the amount of information considered. The benefit from reduced cost of entering evidence may be particularly important for collaborative work. In our task, evidence had already been set up in an easy-to-enter form. But in general the costs of finding and setting up evidence in an easy-to-use form should only be borne only once. An individual user should not have to find and format the same information at different times, and individuals within a collaborative group should not have to bear the cost repeatedly.

Lowering the cost of collaboration: Shared Matrices and Chat

To realize benefits of collaboration, the costs of group processes and accessing information from partners must be low enough that the costs don't outweigh any potential benefits. The Heterogeneous and Solo conditions were similar in performance. This suggests that the process cost of collaboration was relatively low, and did not outweigh the benefits of exposure to mixed opinions.

As suggested in prior work, we used computer-mediated communication, as chat-like tools tend to reduce costs of interaction over face-to-face. Our group-work innovation was ability to view each member's matrix. This provides a highly structured, task-based representation, the same for

each user, which might communicate task-relevant information efficiently. By providing summarized "bottom line" assessment of alternatives, it might help identify differences of opinion and hence support debiasing. In fact, conversations in Chat frequently referred to the bottom-line belief or support in a partner's matrix

CACHE allows participants to read and to add evidence to their own matrix, from a partner's. We did not measure when users are looking at a partner's matrix, but we have one measure of using the partner's evidence. We have logs of interacting with CACHE for two homogeneous and three heterogeneous groups. These give us information about what evidence users read or added. In Blocks 2 and 3, the users in an interacting group each had 10 unique pieces of evidence. Other users could only gain access to these pieces of evidence by looking in their partners' read-only matrices (or the ticker when the partner did something with this evidence). We counted the number of times each user read or added one of these pieces of evidence to their own matrix.

Accessing this information through the partner varied greatly by group and individual. One heterogeneous triad read evidence available only through their partner 31 times, added evidence to their own matrix 22 times, and operated on 21 unique pieces of evidence. Interestingly, this group talked very little and late over chat, and had only 4 turns discussing the problem, all proposing a summary conclusion. One user in a homogeneous group read such evidence 9 times, added 2 pieces to their own matrix, touching 6 unique pieces of evidence. Of the remaining 9 users, 5 never accessed evidence through their partners' matrices and the remaining 4 touched 1 or 2 pieces.

Two points are of interest. First, group strategies varied dramatically. Second, for 4 of the 5 groups with records, the primary way CACHE supported interaction was by allowing viewing the organized information in a partner's matrix, rather than accessing unique information through the partner.

Supporting Use of Negative Evidence.

Teaching the ACH method and providing the CACHE tool supported use of disconfirmatory evidence. Negative evidence, which provides evidence against an hypotheses, is often more difficult to reason with than positive. Users in this experiment were equally likely to include evidence designed to disconfirm (68.5% of the 16 items) as to confirm (69.1% of the 33 items). The evidence structure of this problem is complex, and a confirmation bias could mean reluctance to include disconfirming, negative information, or reluctance to include positive information favoring an alternative hypothesis. Users showed no reluctance to include negative evidence and no [significant] bias for including negative evidence for one hypothesis over another. Apparently, teaching people the ACH method and supporting this with CACHE was sufficient to overcome a bias against negative evidence.

Rich Tool-kit

The CACHE task environment is rich enough for different groups and users to discover alternative strategies. Groups varied in how much they used Chat and in how much they use partner

matrices (to get evidence from). The group which Chatted least pulled evidence from partner matrices the most. Individuals differed in how they used their matrix. We assumed that users would read evidence, evaluate its relevance, and if relevant include in their matrix. Several users, however, included almost everything, and then dropped or gave low relevance to unimportant information. This may indeed have been a very efficient method for systematically reviewing a large amount of evidence.

DISCUSSION

Summary of Findings on Group Process

This study provided computer support to reduce costs of managing the process of collaboration and realize benefits of group diversity. In particular, we asked whether collaboration might improve information coverage and reduce the cognitive bias resulting from overweighting of initial belief and undervaluing of later information.

The task used in our study models a late stage of the analysis of a complex case, when three major alternative hypotheses have already been selected and a large body of evidence relating to these hypotheses has been identified. The evaluation study conducted on the earlier ACH0 interface (Pirolli et al., 2004) had focused on the work done by individual analysts at an earlier stage.

The task in the current study requires a group of analysts to assimilate a substantial amount of background information, including an initial position or hypothesis. This served as the prior belief, which users revised in light of new evidence. CACHE provided users both tools supporting the coordination of evidence and hypotheses at the individual level (the ACH matrix), and tools for sharing information at the group level (sharable partner matrices, evidence, reports on partner actions). Our central question was whether collaborating groups thus supported would do as well as, or even exceed, independent individuals in gathering information and reaching an unbiased final judgment. In particular, when group members have different initial beliefs, this might enable mutual debiasing, and allow individuals in a heterogeneous group to provide a balanced use of evidence, independent of which information they were given first. We did find a significant effect of condition on degree of final bias, even with our small number of groups. Heterogeneous Groups and nominal groups were less biased than the Homogenous Groups. Interpreting this pattern, we suggest that CACHE reduced the overhead of group coordination sufficiently to yield no net process cost for the Heterogeneous Group. Since we did not include groups working without CACHE, any causal claims are only suggestive, of course.

We collected a large amount of information about the judgment process and this enables us to begin identifying which particular processes or judgments were most prone to bias and where any

such bias was most moderated by condition. In addition to effects of condition on bias in the final, user-entered beliefs, we looked for effects on:

a. Amounts of different types of evidence included in the matrix,
b. The summed value of degree of consistency or inconsistency of this evidence,
c. The summed importance of evidence
d. A support measure integrating the importance of each piece of evidence with its confirmatory or disconfirmatory relation to the hypotheses.

We found significant evidence of bias from initial belief on the amount of evidence included in the matrix, particularly positive evidence, on the importance given to this evidence, particularly the positive evidence, and on the values relating evidence to hypotheses. This reveals a form of confirmation bias, that is, the tendency of a judge to look for (and weighting more) confirmatory evidence (Kahneman, Slovic, and Tversky, 1982). However, we only found significant effects of condition on the importance given to positive evidence. Thus, the point at which effects of group structure can most reliably be measured is on evidence importance. Individuals in a Heterogeneous Group and those working alone (Solo group) did not over-weight evidence supporting their favored hypothesis as much as did individuals in a Homogeneous group. Because sensitivity of various measures differs and because we had a small number of groups, our conclusions are certainly suggestive rather than conclusive.

Method for Studying Technology Support for Group Process

Beyond this initial study, we have a promising method for providing detailed information about how group process affects cognitive judgments performed under uncertainty. A key feature of our method is creating and measuring initial bias at an individual level, so debiasing can be measured relatively precisely. A second feature is measuring belief at several points in time. A third feature is use of a problem where the normative answer is roughly equal credibility for each hypothesis. We believe this makes it easier to measure bias in final judgments than would use of an unambiguous problem where the weight of the evidence strongly favored one possibility. Since we are concerned with measuring effects of condition on bias, we need a large, controllable bias to work with. A fourth feature is the effort to differentiate users in terms of role, as well as initial degree of commitment to one or another hypothesis. We hoped to simulate in the lab a characteristic of most work-based teams: individuals have complementary roles, contributing different information and skills. A fifth, unexpected feature is the mix of constraint and openness in this task. The combination of tool-plus-task is constrained enough to give controlled and comparable measures but also rich enough to provoke a variety of unanticipated solution strategies. A sixth characteristic is requiring users to work on each batch of evidence for several minutes before communicating with their partners.

Despite its promise, there are several directions for improvement to the method. First, we tested so much that our measurement disrupted users' work on the task. Even recording belief at the end of each block, which we had thought would be largely part of their normal process, was experienced as distracting or annoying, and it was hard to get users to comply in a timely way. Users were working hard, on their own agenda and did not want to be interrupted. Second, we did not provide explicit instruction for group work in the way the task was presented. Although we suggested different roles, we did not provide any guidance about how these roles might be coordinated. We did not train users on how, specifically, they might use CACHE tools to facilitate the group process. For example, users in a group might search for different types of evidence. Third, although we used our and other's measures of strength of support of the evidence, a more precisely calibrated measure could be developed. Fifth, we tried to create time pressure as a method of convincing users that it was worth their time collaborating (while still asking users to look at each evidence-set on their own). Some users wanted only to collaborate after digesting all the information on their own. Much remains to be explored about how timing of individual and collaborative work influences quality of decision.

A fine-grained assessment method is important because it can tell us which processes are most subject to biases in general, how tools might aid debiasing, and how tools might best focus a debiasing effect of groups. Further, if materials and procedures can be standardized, it would allow comparison of the benefit provided by different combinations of group structure and tool support, to the extent that debiasing can be measured as change from an initial, common bias. The definition of a reference task for investigating collaborative intelligence analysis, as a form of complex decision-making, would facilitate CSCW researchers to focus on shared problems, compare results, identify better design solutions, and improve the measured quality of earlier solutions (Whittaker et al., 2000).

Summary of Results on CACHE Support

When supported by CACHE, users in the Heterogeneous Condition preformed similarly to those working individually. Lack of a difference, here, is noteworthy, because of the various studies, which have found net harm to performance when people are asked to work in groups (e.g., see Steiner, 1972; Kerr and Tindale, 2004). Because many tasks must be done collaboratively due to the sheer size of task, simply equating group with solo performance is an important first step. We believe CACHE reduced the cost of sharing information sufficiently to allow benefits of diverse opinions to balance the process costs of working in a group.

Several other findings about the level of performance here are striking. Users made widespread and extensive use of negative evidence. They were able to incorporate and made use of large amounts of information in a short time. They found the working in CACHE and sharing evidence windows generally helpful, easy to learn, and easy to use. Their solution strategies drew our

attention to many, specific challenges in handling the complexity of information needed for this type of task, in a collaborative manner.

Ideas for Further Development of CACHE

For this user-task combination, heterogeneous groups had their greatest benefit in debiasing the weight placed on included, positive evidence. This may be a particularly informative point for technology-based support in a collaborative system. System support could focus on comparing weights among group members with differing views, localizing evidence with large importance differences, and bring these differences into the group's attention. System support could also focus on comparing weighted and unweighted judgments for each individual user. In fact, if weighting evidence importance is indeed a point very prone to bias, it might be useful to for the system to calculate and display a summary judgment based on the unweighed and the weighed sum of the evidence. Comparing these might help a user identify and compensate for bias.

In addition to trying to debias judgments given a particular group composition, CACHE-like tools could also make recommendations about who should be grouped together into a collaborative group. After individuals have worked alone and formed an initial opinion, groups could be formed based in part on diversity in initial viewpoint (e.g., see method used by Shulz-Hardt et al. (2000) with face-to-face groups).

Collaboration multiplies the information load placed on users, if they attempt to integrate and make use of the information generated by their partners. In the case of CACHE, this is apparent in the large number of windows supporting collaboration, in particular, partner matrices and the ticker window. Designing an interface which selects or integrates key aspects of the partners' work may be helpful. A first step in this direction would be to allow flexible reorganization of matrices to selectively align information by topic relevance or degree of disagreement. In addition, the need for optimizing display of an individual's work is amplified when one must use work from any individuals. Many of our users commented on the need for more flexible control of how their own matrix is displayed.

Conclusion

There is a great need for coordinating multiple knowledge-workers with each other and with large amounts of information. This coordination must be flexible with respect to the dynamics of collaboration, allowing people to work simultaneously on a problem and also to work at different times, using updated information from partners as it arrives. It must support not merely accessing the right information, but incorporating this information into complex decisions.

The judgments and decisions should be based on a broad coverage of the evidence available and should be influenced by as much relevant information as possible, using as normative a means of integration as possible. Judgments and decisions should not be influenced by irrelevant factors, such

as presentation order of evidence, anchoring to prior beliefs, and confirmation bias. CACHE provides a broader array of tools than exercised in the present study. However, even the set in play here show considerable promise for supporting individual judgment and allowing diverse groups to benefit from their diversity. Having shown that the structured use of CACHE can make group processes and outcomes comparable to solo analysis, and having identified additional room for improvement to CACHE, the next step is to determine if collaboration can be further improved to produce superior information coverage and more accurate judgments than solitary analysis.

REFERENCES

Benbasat, I. and Lim, J., 2000. Information technology support for debiasing group judgments: an empirical evaluation. *Organizational Behavior and Human Decision Processes*, 83, 167–183.

Clark, H. H. (1996). *Using language*. Cambridge: Cambridge University Press.

Cummings, J.N. (2004) Work Groups, Structural Diversity, and Knowledge Sharing in a Global Organization. *Management Science*. 50(3), p. 352-364.

Dennis, A. R. & Gallupe, R. B. (1993). A history of group support systems empirical research: lessons learned and future directions. In L. M. Jessup & J. S. Valacich, Eds. *Group Support Systems: New Perspectives*, pp. 59-77. New York: Macmillan.

Dennis, A. R. (1996). Information exchange and use in group decision making: You can lead a group to information, but you can't make it think. *MIS Quarterly*, 20, 433–457.

DeSanctis, G. & Gallupe, R. B. (1987). A foundation for the study of group decision support systems. *Management Science*, 33, pp. 589-609.

Fjermestad, J. An analysis of communication mode in group support systems research, *Decision Support Systems*, 37, (2004) 239-263.

Hackman, J.R. and Kaplan, R.E. Interventions into group process: An approach to improving the effectiveness of groups. *Decision Sciences*, 5 (1974), pp. 459-480.

Heuer, J.R. (1999). *The psychology of intelligence analysis.* Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.

Hollingshead, A. B. (1996). The rank order effect in group decision making. *Organizational Behavior and Human Decision Processes*, 68, 181–193.

Jones, M.D. (1998). *The thinker's toolkit.* NY: Three Rivers Press.

Kent, S. (1949). *Strategic intelligence*. Princeton, NJ: Princeton University Press.

Kerr, N.L. and Tindale (2004). R.S. Group Performance and Decision Making, *Annual Review of Psychology*. 55, 623-655.

Kraut, R., Applying social psychological theory to the problems of group work, in *HCI Models, Theories and Frameworks: Toward A Multidisciplinary Science*, J.M. Carroll, Editor. 2003, Morgan Kaufman: New York. p. 325-356.

Lim, L. and Benbasat, I. The debiasing role of group support systems: an experimental investigation of the representativeness bias, *Int. J. Human-Computer Studies* (1997) 47, 453-471.

McGrath, J., & Hollingshead, A. (1994). *Groups interacting with technology: Ideas, evidence, issues, and an agenda.* Thousand Oaks, CA: Sage.

National Commission on Terrorist Attacks Upon the United States. (2004). *The 9/11 commission report*. New York: Norton.

Nunamaker, J. F., Jr., Dennis, A. R., Valacich, J. S., Vogel, D. R., & George, J. F. (1991). Electronic meeting systems to support group work. *Communications of the ACM*, 34, (7) 40–61.

Parks, C. D., & Cowlin, R. A. (1996). Acceptance of uncommon information into group discussion when that information is or is not demonstrable. *Organizational Behavior and Human Decision* References *Process*, 66, 307-3 15.Cheikes, B. A., Brown, M. J., Lehner, P. E., & Alderman, L. (2004). *Confirmation bias in complex analyses* (No. Technical Report No. MTR 04B0000017). Bedford, MA: MITRE.

Heuer, R. J. (1999). *Psychology of intelligence analysis*. Washington, D.C.: Center for the Study of Intelligence.

Jones, M. D. (1995). *The thinker's toolkit*. New York: Random House.

Krizan, L. (1999). *Intelligence essentials for everyone* (No. Occasional Paper Number Six). Washington, D.C.: Joint Military Intelligence College.

Militello, L. G., & Hutton, R. J. B. (1998). Applied cognitive task analysis (acta): A practitioner's toolkit for understanding cognitive task demands. *Ergonomics, 41*(11), 1618-1641.

Militello, L. G., Hutton, R. J. B., Pliske, R. M., Knight, B. J., & G.Klein. (1997). *Applied cognitive task analysis (acta) methodology* (No. Tech. Rep. No. NPRDC-TN-98-4). Fairborn, OH: Klein Associates.

Scholtz, J. (2004). *Analysis of competing hypotheses evaluation (parc)* (No. Unpublished Report). Gaithersburg, MD: National Institute of Standards and Technology.

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science, 185*, 1124-1131.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology, 12*, 129-140.

Schultz-Hart, S., Frey, D., Lüthgens, C. & Moscovici, S. (2000). Biased Information Search in Group Decision Making. *Journal of Personality and Social Psychology*, 78(4), 655-669.

Stasser, G & Titus, W. (2003). Hidden profiles: A brief history. *Psychological Inquiry*, 14 (3-4), 302-311.

Stasser, G. & Titus, W. (1985). Pooling of unshared information in group decision making: Biased information sampling during discussion. *Journal of Personality and Social Psychology*, 48, 1467-1478.

Stasser, G., and Stewart, D. (1992). Discovery of hidden profiles by decision- making groups: Solving a problem versus making a judgment. *Journal of Personality and Social Psychology*, 63, 426–434.

Steiner ID. 1972. *Group Process and Productivity*. New York: Academic

Stewart, D. D., and Stasser, G. (1995). Expert role assignment and information sampling during collective recall and decision making. *Journal of Personality and Social Psychology*, 69, 619–628.

Straus, S.G. & McGrath, J.E. (1994). Does the medium matter? The interaction of task type and technology on group performance and member reactions. *Journal of Applied Psychology*, 79 (1), 87-89.

Tolcott, M.A., Marvin, F.F., & Lehner, P.E. (1989). Expert decisionmaking in evolving situations. *IEEE Transactions on Systems, Man, and Cybernetics, 19(3)*, 606-615.

Tversky, A., & Kahneman, D. (1974). Judgments under uncertainty. Heuristics and biases. *Science, 185*, 1124-1131.

Whittaker, S., Terveen, L., and Nardi, B. A. (2000). Let's Stop Pushing the Envelope and Start Addressing It: A Reference Task Agenda for HCI, *Human-Computer Interaction*, 15 (2&3), 2000, 75-106.