Field Evaluation in the Intelligence and Counterintelligence Context: Workshop Summary

ISBN
978-0-309-15016-3

114 pages
6 x 9
PAPERBACK (2010)

Robert Pool, Rapporteur; Planning Committee on Field Evaluation of Behavioral and Cognitive Sciences-Based Methods and Tools for Intelligence and Counterintelligence; National Research Council

Field Evaluation in the Intelligence and Counterintelligence Context

Workshop Summary

Robert Pool, *Rapporteur*

Planning Committee on Field Evaluation of Behavioral and Cognitive Sciences-Based Methods and Tools for Intelligence and Counterintelligence

Board on Behavioral, Cognitive, and Sensory Sciences

Division of Behavioral and Social Sciences and Education

NATIONAL RESEARCH COUNCIL
OF THE NATIONAL ACADEMIES

THE NATIONAL ACADEMIES PRESS
Washington, D.C.
www.nap.edu

THE NATIONAL ACADEMIES PRESS 500 Fifth Street, N.W. Washington, DC 20001

NOTICE: The project that is the subject of this report was approved by the Governing Board of the National Research Council, whose members are drawn from the councils of the National Academy of Sciences, the National Academy of Engineering, and the Institute of Medicine. The members of the committee responsible for the report were chosen for their special competences and with regard for appropriate balance.

This study was supported by Grant No. HHQI06-08-C-0010 between the National Academy of Sciences and the U.S. Department of Defense. Any opinions, findings, conclusions, or recommendations expressed in this publication are those of the author(s) and do not necessarily reflect the views of the organizations or agencies that provided support for the project.

International Standard Book Number-13: 978-0-309-15016-3
International Standard Book Number-10: 0-309-15016-7

Additional copies of this report are available from the National Academies Press, 500 Fifth Street, N.W., Lockbox 285, Washington, DC 20055; (800) 624-6242 or (202) 334-3313 (in the Washington metropolitan area); Internet, http://www.nap.edu.

Printed in the United States of America

Cover credit: Image in magnifying glass: U.S. Army 1st Lt. Michael Frank (center) of Charlie Company, 1st Battalion, 327th Infantry Regiment, questions a gasoline vendor along the main road next to Rashad, Iraq, during a patrol, May 23, 2006. U.S. Navy photo by Petty Officer 1st Class Jeremy L. Wood.

Suggested citation: National Research Council. (2010). *Field Evaluation in the Intelligence and Counterintelligence Context: Workshop Summary*. Robert Pool, Rapporteur. Planning Committee on Field Evaluation of Behavioral and Cognitive Sciences-Based Methods and Tools for Intelligence and Counterintelligence. Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

THE NATIONAL ACADEMIES

Advisers to the Nation on Science, Engineering, and Medicine

The **National Academy of Sciences** is a private, nonprofit, self-perpetuating society of distinguished scholars engaged in scientific and engineering research, dedicated to the furtherance of science and technology and to their use for the general welfare. Upon the authority of the charter granted to it by the Congress in 1863, the Academy has a mandate that requires it to advise the federal government on scientific and technical matters. Dr. Ralph J. Cicerone is president of the National Academy of Sciences.

The **National Academy of Engineering** was established in 1964, under the charter of the National Academy of Sciences, as a parallel organization of outstanding engineers. It is autonomous in its administration and in the selection of its members, sharing with the National Academy of Sciences the responsibility for advising the federal government. The National Academy of Engineering also sponsors engineering programs aimed at meeting national needs, encourages education and research, and recognizes the superior achievements of engineers. Dr. Charles M. Vest is president of the National Academy of Engineering.

The **Institute of Medicine** was established in 1970 by the National Academy of Sciences to secure the services of eminent members of appropriate professions in the examination of policy matters pertaining to the health of the public. The Institute acts under the responsibility given to the National Academy of Sciences by its congressional charter to be an adviser to the federal government and, upon its own initiative, to identify issues of medical care, research, and education. Dr. Harvey V. Fineberg is president of the Institute of Medicine.

The **National Research Council** was organized by the National Academy of Sciences in 1916 to associate the broad community of science and technology with the Academy's purposes of furthering knowledge and advising the federal government. Functioning in accordance with general policies determined by the Academy, the Council has become the principal operating agency of both the National Academy of Sciences and the National Academy of Engineering in providing services to the government, the public, and the scientific and engineering communities. The Council is administered jointly by both Academies and the Institute of Medicine. Dr. Ralph J. Cicerone and Dr. Charles M. Vest are chair and vice chair, respectively, of the National Research Council.

www.national-academies.org

PLANNING COMMITTEE ON FIELD EVALUATION OF BEHAVIORAL AND COGNITIVE SCIENCES–BASED METHODS AND TOOLS FOR INTELLIGENCE AND COUNTERINTELLIGENCE

Philip E. Rubin (*Chair*), Haskins Laboratories, Yale University
Robert F. Boruch, Graduate School of Education and Statistics, University of Pennsylvania
Robert A. Fein, Department of Psychiatry, McLean Hospital, Harvard Medical School
Jonathan D. Moreno, Department of Medical Ethics and Department of History and Sociology of Science, University of Pennsylvania
Eduardo Salas, Institute for Simulation and Training, University of Central Florida
Neil Thomason, Department of History and Philosophy of Science (retired), University of Melbourne
Carol H. Weiss, Graduate School of Education, Harvard University

Bud Pautler, *Study Director* (through April 2009)
Mary Ellen O'Connell, *Study Director* (from April 2009)
Robert Pool, *Rapporteur*
Renée L. Wilson Gaines, *Senior Program Assistant*

Acknowledgments

This workshop summary is based on the discussion at a workshop convened by the Board on Behavioral, Cognitive, and Sensory Sciences on September 22-23, 2009, and planned by the Committee on Field Evaluation of Behavioral and Cognitive Sciences-Based Methods and Tools for Intelligence and Counterintelligence. The planning committee members identified presenters, organized the agenda, made presentations, and facilitated discussion, although they did not participate in the writing of this report. This summary reflects their diligent efforts, the excellent presentations by other experts at the workshop, and the insightful comments of the many workshop participants.

The workshop was sponsored by the Defense Intelligence Agency and the Office of the Director of National Intelligence. The interest and support of Susan Brandon, chief for research, Behavioral Science Program DEO–Defense CI and HUMINT Center Defense Intelligence Agency, and Steven Rieber, research director, Office of Analytic Integrity and Standards, Office of the Director of National Intelligence, are much appreciated.

This summary has been reviewed in draft form by individuals chosen for their diverse perspectives and technical expertise, in accordance with procedures approved by the Report Review Committee of the National Research Council. The purpose of this independent review is to provide candid and critical comments that will assist the institution in making its published report as sound as possible and to ensure that the report meets institutional standards for objectivity, evidence, and responsiveness to the charge. The review comments and draft manuscript remain confidential

to protect the integrity of the process. We thank the following individuals for their review of this report: Robert A. Fein, Department of Psychiatry, McLean Hospital, Harvard Medical School; Carl W. Ford, Jr., National Intelligence Council Associate, Office of the Director of National Intelligence; Leslie K. Goodyear, Division of Research on Learning in Formal and Informal Settings, National Science Foundation; Elizabeth F. Loftus, Departments of Psychology and Social Behavior and Criminology, Law and Society, University of California, Irvine; and Christian A. Meissner, Departments of Psychology and Criminal Justice, University of Texas, El Paso.

Although the reviewers listed above provided many constructive comments and suggestions, they were not asked to endorse the content of the report, nor did they see the final draft of the report before its release. The review of this report was overseen by John T. Monahan, School of Law, University of Virginia. Appointed by the National Research Council, he was responsible for making certain that an independent examination of this report was carried out in accordance with institutional procedures and that all review comments were carefully considered. Responsibility for the final content of this report rests entirely with the author and the institution.

Contents

1

Introduction

On September 22-23, 2009, the Board on Behavioral, Cognitive, and Sensory Sciences of the National Research Council held a workshop on the field evaluation of behavioral and cognitive sciences–based methods and tools for use in the areas of intelligence and counterintelligence.[1] Broadly speaking, the purpose of the workshop was to discuss the best ways to take methods and tools from behavioral science and apply them to work in intelligence operations. More specifically, the workshop focused on the issue of field evaluation—the testing of these methods and tools in the the context in which they will be used in order to determine if they are effective in real-world settings.

This report is a summary and synthesis of the two days of presentations and discussions that took place during the workshop.[2] The workshop participants included the members of the committee that planned the workshop, along with invited speakers and a number of other participants, including experts from a number of areas related to the behavioral sciences and the intelligence community. The goal of the workshop was not to provide specific recommendations but to offer some insight—in large part through specific examples taken from other fields—into the sorts of issues that surround the area of field evaluations. The discussions

[1]For ease of reading, the phrase "intelligence and counterintelligence" is not repeated throughout the summary. Such terms as "intelligence community" and "intelligence operations" are intended to include both intelligence and counterintelligence.

[2]Presentations from the workshop are available at: http:// nationalacademies.org/bbcss/Field_Evaluation_Workshop_Presentations.html.

covered such ground as the obstacles to field evaluation of behavioral science tools and methods, the importance of field evaluation, and various lessons learned from experience with field evaluation in other areas.

It is important to be specific about the nature of this report, which documents the information presented in the workshop presentations and discussions. Its purpose is to lay out the key ideas that emerged from the workshop and should be viewed as an initial step in examining the research and applying it in specific policy circumstances. The report is confined to the material presented by the workshop speakers and participants. Neither the workshop nor this summary is intended as a comprehensive review of what is known about the topic, although it is a general reflection of the literature. The presentations and discussions were limited by the time available for the workshop. A more comprehensive review and synthesis of relevant research knowledge will have to wait for further development.

This report was prepared by a rapporteur and does not represent findings or recommendations that can be attributed to the planning committee. Indeed, the report summarizes views expressed by workshop participants, and the committee is responsible only for its overall quality and accuracy as a record of what transpired at the workshop. Also, the workshop was not designed to generate consensus conclusions or recommendations but focused instead on the identification of ideas, themes, and considerations that contribute to understanding the current state of field evaluation of behavioral and cognitive sciences–based methods and tools for use in the areas of intelligence and counterintelligence.

To fully appreciate the workshop, the reader needs two important bits of context. The first is the relationship between the behavioral sciences and the intelligence community and, in particular, what the intelligence community has to gain from establishing a close relationship with the community of behavioral scientists. The second is the current urgency to improve the performance and capabilities of the intelligence community.

THE BEHAVIORAL SCIENCES AND THE INTELLIGENCE COMMUNITY

In one of the workshop presentations, David Mandel, a senior defense scientist at Defence Research and Development Canada (DRDC), discussed the ways in which the behavioral sciences can benefit intelligence analysis and why it is important for the intelligence community to build a partnership with the behavioral sciences community.

First, however, Mandel offered a working definition of *behavioral science*: it is science aimed at understanding human behavior in a broad

sense, including both the causes and the consequences of that behavior. As such, it includes a variety of scientific fields, such as psychology, sociology, anthropology, political science, economics, and, on the biological side, the neurosciences. Although traditionally these fields have been seen as separate areas of science, increasingly they have come to overlap and intersect, to the point that behavioral science is more of a continuum than a collection of independent fields.

The intelligence community has long relied on science and technology for insights and techniques, Mandel noted, so one might wonder why it is necessary to talk about the importance of strengthening the relationship between the intelligence community and the broad community of behavioral scientists. One important reason, he said, is that there are a number of factors that tend to weaken the relationship between the two communities and make analysts less likely to take advantage of what the behavioral sciences can offer.

First, Mandel said, there is a natural inclination among most people—including those in the intelligence community—to react poorly to "scholarly verdicts that deal with issues such as the quality of their judgment and decision making, their susceptibility to irrational biases, their use of suboptimal heuristics, and overreliance on nondiagnostic information." Like most people, experts have the sense that they are competent. Psychological research shows that most people believe themselves to be better than average at what they do. Thus, Mandel said, experts are prone to challenge conclusions offered by behavioral scientists with their own knowledge gained from personal experience and, furthermore, to believe that such a challenge is completely legitimate. This is a fundamental problem that behavioral scientists face in making contributions to any practitioner community, Mandel said, "Their research is very easily disregarded on the basis of intuition and common sense."

A second reason that analysts tend to disregard lessons from behavioral science is that it is seen as being "soft" science. Thus its knowledge is considered to be less objective or trustworthy than knowledge generated by the "hard" sciences and technology, such as satellite imaging or electronic eavesdropping. Although that attitude is common in the intelligence community, Mandel cautioned, it is misguided and underestimates both the value and the analytical power of behavioral science. "When someone uses the term 'soft science,' I correct them. I say 'probabilistic science' and [note that] we deal with some very difficult problems."

Third, Mandel said, the relationship between the intelligence community and the behavioral science community is still relatively new, so analysts do not necessarily understand what behavioral science has to offer. Thus, he noted, forums like this workshop are important for explor-

ing ways in which the partnership between the two communities can be developed.

Developing such a partnership is important for a number of reasons, Mandel said. From 1978 to 1986, Richards J. Heuer, Jr., an analyst with the Central Intelligence Agency, wrote a number of articles surveying the cognitive psychology literature, translating it into terms that other analysts could easily understand, and suggesting ways that those research findings could be applied to improve performance in various tasks undertaken by the intelligence community. The articles were later collected in a book, *Psychology of Intelligence Analysis* (Heuer, 1999).

It was a remarkable feat, Mandel said, for one person from outside the field of cognitive psychology not only to effectively interpret a significant portion of the literature in that field but also to come up with recommendations for various procedures based on that literature which would, in many cases, become part of the training and practice of intelligence analysts. But it is because Heuer's accomplishment was so singular that it becomes clear that there should be some mechanism or systematic arrangement for applying insights and knowledge from behavioral science to the field of intelligence analysis. The intelligence community cannot afford to rely on the occasional emergence of an inspired maverick like Heuer to make those connections.

Mandel offered several supporting arguments for this conclusion. The first is an opportunity cost argument: Heuer's work has such a valuable payoff for the intelligence community that maintaining the status quo—with no established mechanisms for applying behavioral science to intelligence analysis—means missing out on many valuable applications that could be expected from a more systematic effort to exploit knowledge from behavioral science.

Second, relying on the occasional maverick is not a good way for the intelligence community to remain current on what is being discovered in the diverse areas of behavioral science. Intelligence analysts have their own full-time jobs; they cannot be expected to also keep up with all the relevant advances in research in behavioral science and determine how to integrate that knowledge into intelligence work.

It is telling, Mandel noted, that no one else has come along since Heuer to continue his work of translating cognitive psychology and other areas of behavioral science into tools for analysis. "In cognitive psychology alone there is at least a quarter century of new research since Heuer published *Psychology of Intelligence Analysis* that is waiting to be exploited by the intelligence community."

Another way in which establishing a connection with the research community can help the intelligence community is with validation, Mandel said. Once knowledge and insights from behavioral science are

used to develop new tools for the intelligence community, it is still necessary to validate them. Simply basing recommendations on scientific research is not the same thing as showing scientifically that those recommendations are effective or testing to see if they could be substantially improved. Even Heuer was unable to do much to validate his recommendations, Mandel noted, and, more generally, this is not something that the intelligence community is particularly well equipped to do.

It is, however, exactly what research scientists are trained to do. Science offers a method for testing which ideas lead to good results and which do not. Thus partnering with the behavioral science community can help the intelligence community zero in on the techniques that work best and avoid those that work poorly or not at all.

In theory, Mandel said, it would be possible for the intelligence community to build its own applied behavioral research capability, but that would draw significant resources away from other operational areas and add an entirely new focus and purpose to the intelligence community's existing tasks. Furthermore, if the intelligence community were to hire behavioral scientists, it would find itself in competition with both academia, with its unparalleled freedoms, and industry, with its lucrative salaries. It makes more sense, Mandel suggested, for the intelligence community to develop partnerships with universities and other institutions that already have the expertise and capability to perform behavioral science research.

A final advantage of partnering with the existing behavioral science community, Mandel said, is the "multiplier effect." By working with scientists in academia, for example, the intelligence community is not only drawing on the knowledge of those subject-matter experts but on all of their contacts. "As a researcher in an R&D [research and development] organization and government," Mandel said, "I am very keen on partnering with academics because I understand that they have the ability to reach back into other areas of academia and connect me with other experts who could be of use." There is a tremendous amount of such leverage that can be achieved by building relationships rather than trying to do everything in-house.

URGENCY

There is a good deal more pressure to perform on the U.S. intelligence community now than at any time in the recent past. There is a major threat that requires accurate intelligence to combat. For evidence one need look no further than the terrorist attacks of September 11, 2001. Although the intelligence community has successfully identified other terror plots before they could come to fruition, this one was missed, and more than

3,000 people died. At the same time, U.S. troops in Iraq and Afghanistan are faced with regular threats, and one important line of defense is the work of intelligence officers there.

As Steven Kleinman, a consultant on intelligence and national security policy, pointed out, one of the key tools in dealing with such threats is HUMINT, or human intelligence. Although there is a great deal that can be learned with SIGINT (signals intelligence) or IMINT (imaging intelligence), much of the work of the intelligence community inevitably relies on such human-centered activities as asking questions and figuring out if someone is lying or predicting what someone will do in a particular situation, and HUMINT is generally acknowledged as the more important element in defeating terrorism and winning wars. And it is in HUMINT that insights and techniques from the behavioral sciences offer the potential of providing new and improved capabilities. Just as dramatically improved satellite imaging technology has greatly increased the capabilities of IMINT, Kleinman said, the hope is that behavioral science can improve the country's HUMINT capacities. "We are in a multifront war," he said. "There are some young men and women, and not so young men and women, who are out there putting their lives on the line." They deserve the best and most accurate intelligence that can be provided, he said.

In a similar vein, Anthony Veney, chief of counterintelligence investigation and functional services at U.S. Central Command, offered a dramatic description of the stakes and expressed his hopes that the scientific community could soon provide some tools and techniques that could make a difference in the field.

First, he said, it is important to understand that the fight in Iraq and Afghanistan is, in essence, an information war. It is less about bullets than about who controls the airwaves, he said, and about who is getting their message across the fastest to the most people with the most credibility.

Second, one of the most important tasks of intelligence officers in Iraq and Afghanistan is to determine who from among all the people with whom they come in contact can be trusted. Here, for example, is the sort of person they must make a decision about: "He is a tribal elder today. He is an intelligence source tomorrow. He is a drug trafficker on Wednesday. He doesn't care how he makes his money; he is just trying to make money." It is in this person's best interest to provide everybody with information about what is going on, be it the Iranians, the Russians, the Taliban, the Chinese, or even some American who wants some information. So how much can he be trusted? How can it be detected if he starts to lie? These are potentially questions of life and death.

But these are questions that must be answered without a large intelligence infrastructure. Such a structure does not provide enough flexibil-

ity or time to react. So, Veney said, "We have infantry officers relying on translators for information. Do I raid this house? Do I raid that house?"

At the same time, many of the support services that historically were provided by army members are now provided by locals. So now it is important to be able to tell—with most of the people involved speaking a foreign language—which people can be trusted enough to be let onto the base on a regular basis. "People are provided access that 40 years ago, 50 years ago we would have never given to our facilities. Those are the people we are trying to discern what is it that they are doing. . . . Are they being honest?"

So, Veney said, there is an urgent need for devices that perform such functions as accurately detecting when someone is lying. "What I am asking for is for people to hurry up because we don't have time for years and years of research," he said. "Give me something I can use on the battlefield. It can't be as big as an MRI machine—I can't move that." It also needs to be simple enough that a soldier can be trained to use it in 24 hours. "If a dog smells fear, why can't I have a pen that can tell me if somebody is lying? That is what I am asking for. I am losing friends out there."

This pressure to save lives is a major driving force behind the current interest in applying behavioral science to intelligence, noted Robert Fein, a forensic psychologist at Harvard Medical School and a planning committee member, but it is a pressure that must be resisted to a certain degree if the science is to be done correctly. As discussed at various points in the workshop, a sense of urgency can lead to techniques and devices being adopted before they have been carefully evaluated, and this in turn can lead to reliance on methods that are ineffective or are less effective than available alternatives. Indeed, the purpose of field evaluation is to avoid such situations by determining what works and what does not. But effective field evaluations take time and thus can come into conflict with an urgent sense that something needs to be done *now*, if not sooner.

And, indeed, the issue of finding the right balance between the urgency and the need for field evaluation was one of the themes underlying much of the discussion throughout the workshop. It is important to provide the men and women of the intelligence community with new and improved tools to help them do their jobs, but it is equally important to take the time to make sure that those tools actually work.

2

Behavioral Tools and Techniques

In what ways might particular tools and techniques from the behavioral sciences assist the intelligence and counterintelligence community? A variety of devices and approaches derived from the behavioral sciences have been suggested for use or have already been used by the intelligence community. In the workshop's first session, speakers described several of these, with a particular emphasis on how the techniques have been evaluated in the field. As Robert Fein put it, "Our spirit here is to move forward, to figure out what kinds of new ideas, approaches, old ideas might be useful to defense and intelligence communities as they seek to fulfill what are often very difficult and sometimes awesome responsibilities."

To that end the speakers provided case studies of various technologies with potential application to the intelligence field. One common thread among all of these disparate techniques, a point made throughout the workshop, is that none of them has been subjected to a careful field evaluation.

DECEPTION DETECTION

People in the military, in law enforcement, and in the intelligence community regularly deal with people who deceive them. These people may be working for or sympathize with an adversary, they may have done something they are trying to hide, or they may simply have their own personal reasons for not telling the truth. But no matter the reasons,

an important task for anyone gathering information in these arenas is to be able to detect deception. In Iraq or Afghanistan, for example, soldiers on the front line often must decide whether a particular local person is telling the truth about a cache of explosives or an impending attack. And since research has shown that most individuals detect deception at a rate that is little better than random chance, it would be useful to have a way to improve the odds. Because of this need, a number of devices and methods have been developed that purport to detect deception. Two in particular were described at the workshop: voice stress technologies and the Preliminary Credibility Assessment Screening System.

Voice Stress Technologies

Of the various devices that have been developed to help detect lies and deception, a great many fall in the category of voice stress technologies. Philip Rubin, chief executive officer of Haskins Laboratories and chair of the Board on Behavioral, Cognitive, and Sensory Sciences at the National Research Council, offered a brief overview of these technologies and of how well they have performed on objective tests.

The basic idea behind all of these technologies, he explained, is that a person who answers a question deceptively will feel a heightened degree of stress, and that stress will cause a change in voice characteristics that can be detected by a careful analysis of the voice. The change in the voice may not be audible to the human ear, but the claim is that it can be ascertained accurately and reliably by using signal-processing techniques.

More specifically, many of the voice stress technologies are based on the assumption that microtremors—vibrations of such a low frequency that they cannot be detected by the human ear—are normally present in human speech but that when a person is stressed, the microtremors are suppressed. Thus by monitoring the microtremors and noting when they disappear, it should be possible to determine when a person is speaking under stress—and presumably lying or otherwise trying to deceive.

A number of different voice stress technologies have been manufactured and marketed, most of them to law enforcement agencies, but some also to insurance fraud investigators and to various intelligence organizations, including a number in the U.S. Department of Defense. One of the earliest products was the Psychological Stress Evaluator from Dektor Corporation. Originally developed in 1971, it has gone through numerous modifications, and a version is still being sold today. Other voice stress technologies include the Digital Voice Stress Analyzer from the Baker Group, the Computer Voice Stress Analyzer from the National Institute for Truth Verification, the Lantern Pro from Diogenes, and the Vericator from Nemesysco.

Over the years, these technologies have been tested by various researchers in various ways, and Rubin described a 2009 review of these studies that was carried out by Sujeeta Bhatt and Susan Brandon of the Defense Intelligence Agency (Bhatt and Brandon, 2009). After examining two dozen studies conducted over 30 years, the researchers concluded that the various voice stress technologies were performing, in general, at a level no better than chance—a person flipping a coin would be equally good at detecting deception. In short, there was no evidence for the validity or the reliability of voice stress analysis for the detection of deception in individuals.

Furthermore, Rubin said, not only is there no evidence that voice stress technologies are effective in detecting stress, but also the hypothesis underlying their use has been shown to be false. If indeed there are microtremors in the voice, then they must result from tremors in some part of the vocal tract—the larynx, perhaps, or the supralaryngeal vocal tract, which is everything above the larynx, including the oral and nasal cavities. Using a technique called electromyography to measure the electrical signals of muscle activities, physiologists have found that there are indeed microtremors of the correct frequency—about 8 to 12 hertz—in some muscles, including those of the arm. So it would seem reasonable to think that there might also be such microtremors in the vocal tract, which would produce microtremors in the voice. However, research has found no such microtremors, either in the muscles of the vocal tract or in the voice itself. So the basic idea underlying voice stress technologies—that stress causes the normal microtremors in the voice to be suppressed—is not supported by the evidence.

Rubin did not claim that voice stress technologies do not work, only that there has been extensive testing with very little evidence that such technologies do work. It is possible that some of the technologies do work under certain conditions and in certain circumstances, but if that is so, more careful testing will be needed to determine what those conditions and circumstances are. And only when such testing has been carried out and the appropriate conditions and circumstances identified will it make sense to carry out field evaluations of such technologies. At this point, voice stress technologies are not ready for field evaluation.

For the most part, Rubin said, the intelligence community has now stayed away from voice stress technologies mainly because of the absence of any evidence supporting their accuracy. But the law enforcement community has taken a difference approach. Despite the lack of evidence that the various voice stress technologies work, and despite the absence of any field evaluations of them, the technologies have been put to work by a number of law enforcement agencies around the country and around the world. It is not difficult to understand the reasons, Rubin said. The

devices are inexpensive. They are small and do not require that sensors be attached to the person being questioned; indeed, they can even be used in recorded sessions. And they require much less training to operate than a polygraph.

Many people in law enforcement believe that the voice stress technologies do work; even among those who are convinced that the results of the technologies are unreliable, many still believe that the devices can be useful in interrogations. They contend that simply questioning a person with such a device present can, if the person believes that it can tell the difference between the truth and a lie, induce that person to tell the truth.

Preliminary Credibility Assessment Screening System

With the reliability of voice stress technologies called into question, the intelligence community needed another way to screen for deception. Donald Krapohl, special assistant to the director of the Defense Academy for Credibility Assessment (DACA), described to the workshop audience how, several years ago, the Pentagon asked DACA for a summary of the research on voice stress technologies. DACA, which is part of the Defense Intelligence Agency in the Department of Defense, provided a review of what was known about voice stress analysis, and, as Krapohl put it, "it was rather scary to them, and they decided to pull those technologies back."

The need for deception detection remained, however, and DACA's headquarters organization, the Counterintelligence Field Activity (CIFA),[1] was given the job of finding a new technology that would do the same job that voice stress technologies were supposed to perform but with significantly more accuracy. There were a number of requirements in order for a device to be effective in the field: it had to have low training requirements, as it would be used by soldiers on the front line rather than interrogation specialists; ideally it would require no more than a week of training. It needed to be highly portable and easy to use for the average soldier. It needed to be rugged, as inevitably it would be dropped, get wet, and get dirty.

And it had to be a deception test, not a recognition test. That is, instead of recognizing when someone knows something that they are trying to hide—the so-called guilty knowledge test—it should be able to detect when someone was giving a deceptive answer to a direct question. There is a great deal of research concerning the guilty knowledge

[1]CIFA was shut down in 2008 and its responsibilities were taken over by a new agency, the Defense Counterintelligence and Human Intelligence Center.

test, Krapohl explained, but the test is not particularly useful in the field because the interviewers must know something about the "ground truth." Deception tests, by contrast, are not as well understood by the scientific community, but they are far more useful in the field, where interviewers may not know the ground truth.

The final requirement for the device was that it needed to be relatively accurate as an initial screening tool. It was never intended to provide a final answer of whether someone was telling the truth. Its purpose instead was to provide a sort of triage: when soldiers in the field question someone who claims to have some information, they need to weed out those who are lying. The ones who are not weeded out at this initial stage would be questioned further and in more detail. There are polygraph examiners who can perform extensive examinations, Krapohl explained, but their numbers are limited. "So if you could use a screening tool up front to decide who gets the interview, who gets the interrogation, who gets the polygraph examination, the commanders thought that would be very useful," he said. "It was not designed to be a standalone tool. It was designed only as an initial assessment."

The contract to develop such a device was given to the Applied Physics Laboratory at Johns Hopkins University along with Lafayette Instruments. The first prototype of the instrument, called the Preliminary Credibility Assessment Screening System, or PCASS, was finished in January 2006, and it was delivered to DACA for validation.

The PCASS consists of three sensors connected to a personal digital assistant. Two of the sensors are electrodermal sensors, which measure the electrical conductivity of the skin, and one is a photoplethysmograph, which is attached to a finger and used to measure changes in blood flow. The signals from the sensors are fed through an analog-to-digital converter and sent to the personal digital assistant for analysis.

The PCASS is used very much like a polygraph, with an interview phase in which the person being tested is asked a series of questions about such things as personal health, followed by a review of the test questions, then the asking of the test questions, which are designed to be answered with a yes or a no. Where PCASS differs from a polygraph is in how the test results are presented. Unlike a polygraph, which delivers a collection of data records that must be interpreted by someone trained in polygraphy, the PCASS device flashes the words "red," "yellow," or "green" on a screen. "Red" indicates that the person had significant physiological reactions when asked significant questions, indicating that the person may have provided deceptive answers. "Green" indicates no significant physiological responses to the relevant questions, and "yellow" appears when the test was inconclusive.

One of the key facts about PCASS is that it was designed specifically

to detect deception, which made it possible, Krapohl said, to create an algorithm that considers all of the response data and provides a straightforward answer to the question of whether a person is being deceptive: yes, no, or maybe. It does not provide nearly as much information as a polygraph can, but that is not its purpose.

Krapohl described three tests that have been done to assess the accuracy of PCASS, two of them carried out by DACA and one by Battelle. The DACA tests were simulated bombing scenarios in which U.S. Army trainees at Ft. Jackson in South Carolina were randomly assigned to two groups, bombers and innocents. The bombers were instructed to take a simulated bomb and place it near a vehicle in a simulated village called Dragon City on the Ft. Jackson grounds; the innocents were told to walk to a particular area and stay there for 15 minutes. Afterward both innocents and bombers were given a PCASS examination by mid-level enlisted soldiers who had been recently trained by DACA on using PCASS.

The Battelle test was quite different. It was carried out in Columbus, Ohio, by people recruited from the community. The scenario involved the theft of a diamond ring from an office, and the PCASS examinations were carried out by retired law enforcement officers who had been recently trained in the use of PCASS.

The results of the three tests were significantly better than chance. Of the 109 guilty people in the three studies, the PCASS operators identified 75 of them as guilty, 11 of them as innocent, and the results for 23 were inconclusive. Of the 117 innocent people, the operators identified 78 of them as innocent, 19 as guilty, and the results for 20 were inconclusive.

If the inconclusives are considered to be errors, the operators were correct about 68 percent of the time. However, Krapohl noted, in practice the inconclusive results are not exactly errors because in the field a person who scores inconclusive will be given the test again in a day or two. If the inconclusive scores are not included in the results, the operators scored nearly 84 percent correct. The accuracy rates were highest in the Battelle test, whose operators got nearly 92 percent of the identifications right when the inconclusive scores are not considered.

It is not clear why the Battelle group scored higher, Krapohl said. The two scenarios were quite different, with one involving the planting of a bomb and the other the theft of a ring, and that may be part of the explanation. Perhaps the retired law enforcement officers who performed the tests had interrogation experience that allowed them to perform better even with a device they had just learned to use, or perhaps the civilians did not do as good a job as the Army soldiers in fooling the test. Whatever the reason for the difference, Krapohl said, the three tests indicate that PCASS does a good job of detecting deception in these created sce-

narios even when operated by people with relatively little experience in its use.

The technique is now being used by U.S. forces in a number of arenas, said Anthony Veney. Those arenas include Iraq and Afghanistan as well as Colombia, where PCASS is being used in counter-drug operations, he said.

As designed, the main function of the device has been as an initial assessment of whether someone is being deceptive, Veney said. "We use this as a triage tool. We needed to be able to quickly tell if an individual on the battlefield was telling us the truth. Once the person is back in an interrogation center, we have a very expert corps of interrogators that will have at him. We can use other tools like polygraphs to determine credibility."[2]

PCASS allows soldiers who have been trained quickly in the technique to determine whether a person needs to be questioned further or to decide whether a person should be allowed to work on a U.S. forward operating base or in a U.S. installation in a battle zone, Veney explained. For example, for a source reporting information that he says he has collected, if he comes back red on a PCASS test, it is reasonable to assume that the source is not being honest and is not providing reliable information, so no further attention is paid to what he has to say. Similarly, if someone is requesting employment on or access to a base but comes up red on the test, he is sent away. However, Veney said, coming up green is not considered proof of trustworthiness. A person who comes up green on one test will continue to be tested over time.

In short, the main use for PCASS is on the front lines where soldiers need help in determining who seems trustworthy and who seems to have something to hide. But the technique is not assumed to give a definite answer, only a conditional one.

Because PCASS is used on the front lines, it has never been field tested, Veney explained. "This is way too dangerous on the battlefield, to have scientists roaming around doing additional research." Still, it has proved its value in various ways, he said. In a recent operation in Iraq, for example, it allowed U.S. forces to identify a number of individuals who were working for foreign intelligence services and others who were working for violent extremist organizations. "It has been a godsend on the battlefield," Veney said. In Colombia, PCASS made it possible to determine that several people who had claimed to the Colombian government

[2]A National Research Council report (2003) questioned the value of polygraph tests, concluding that the scientific evidence for the accuracy of the polygraph as a screening or deception diction tool is limited.

that they belonged to FARC, the Revolutionary Armed Forces of Colombia, actually did not belong to it.

Still, Krapohl said, there is more work to be done. The group at DACA thinks, for example, that by taking advantage of some of the state-of-the-art technologies for deception detection, it should be possible to develop more accurate versions of PCASS. In particular, by using the so-called directed lie approach—in which those being questioned are instructed to provide false answers to certain comparison questions—it should be possible to get greater standardization and less intrusiveness, he said.

Still, the issue of field evaluation remains, Krapohl said. Although the technique has been tested in the laboratory, there are no data on its performance in the field. "Doing validation studies of the credibility assessment technology in a war zone has a number of problems that we have not been able to figure out," he said. Nonetheless, DACA researchers would like to come up with ideas for how PCASS and other credibility assessment technologies might be evaluated in the field.

In later discussions at the workshop, it became clear that a number of participants had serious doubts about the effectiveness of PCASS in the field, despite the fact that it is in widespread use and, as Veney noted, popular among at least some of the troops in the field. "Everybody in this room knows that there are real limitations to it," Fein said. "I think we can do better than put something out there that has such limitations." And Brandon commented that "if we were doing really good field validation with the PCASS" then it might well become obvious that other, less expensive methods could do at least as good a job as PCASS at detecting deception. There are a number of important questions concerning the validity and reliability of PCASS that can be addressed only by field evaluation, and until such validation is done, the troops in the field are relying on what is essentially an unproved technology.[3]

PREDICTION METHODS

One of the most common tasks given to intelligence analysts is to predict the future. They may be asked, for example, to forecast the chances

[3]See Bhatt and Brandon (2008), which examines thoroughly the unresolved issues and concerns surrounding PCASS and, in particular, the problems that arise from using the system without field evaluation. In particular, the authors note that there is no evidence for the validity of PCASS in the field and that there are several reasons why the success in the DACA and Battelle experiments might not translate into success in the field. They note that other means of screening are available in the field, such as human judgment, and there is no evidence that screening with PCASS is more effective than the alternatives. And they note that reliance on a technology such as PCASS often leads people to suspend their own judgment and defer to the technology, even if their judgment might be superior.

that one country will invade another or to predict whether a dictator will be overthrown. Peering into the future in this way demands a great deal of information about the present—much of which may be uncertain or simply not available—along with an understanding of the psychology of individuals, the dynamics of groups, and the inner workings of government bodies.

As Neil Thomason of the University of Melbourne noted in his workshop presentation, the standard technique in the intelligence community for making such predictions has been what might be called the expert judgment model: "You know the material, you talk with your colleagues, you think it over a lot, and you write up your final thoughts." It is an approach that depends on individual analysts applying their knowledge, experience, and judgment to come up with the best predictions they can.

But how good are the predictions made by this expert judgment approach? As a partial answer, Thomason offered some data concerning the difficulties experts have in accurately predicting U.S. Supreme Court decisions.

Such predictions are extremely difficult to get right, he noted. There are nine Supreme Court justices, each with his or her idiosyncrasies, and the cases are heard against a huge background of legal precedent and, often, conflicting political agendas. Ruger and colleagues (2004) compared the predictions of 2002 Supreme Court decisions made by legal experts with those made by a crude flow chart, generated more or less mechanically and without any understanding of the legal issues involved. The issue for each case was a simple yes-or-no question: Will the Supreme Court reverse the ruling of the lower court?

The experts were from major law schools or appellate attorneys, and they made predictions only on cases in their areas of expertise. The study found that they were right about 59 percent of the time—better than flipping a coin, but not by much. The crude flow chart did much better, getting the right answer 75 percent of the time.

Similarly, Thomason said, a meta-analysis by Grove and colleagues (2000) examined about 140 studies pitting expert judgment against actuarial models, some of them very crude. Of these studies, the experts outperformed the actuarial models in only 8, the models outperformed the experts in 65, and the performance of the experts and the models was about the same in the remaining 63.

None of this shows that analysts in the intelligence community could be outperformed by predictive models, Thomason said, but it does suggest the possibility that such models can be used to improve expert judgments. And, indeed, over the past several decades, the intelligence community occasionally has used various models and approaches to help improve

predictions. The speakers at the workshop discussed two approaches in particular: structured-thinking techniques, also known as structured analytic techniques, and Bayesian analysis.

Alternative Competing Hypotheses

The basic idea behind structured-thinking techniques, Thomason explained, is to help experts structure their thinking so that various biases are alleviated or even avoided altogether. There is a well-known, well-established psychological literature on such biases that informs the techniques.

In intelligence circles, the best known and most commonly used structured-thinking technique is called Analysis of Competing Hypotheses, or ACH. It was developed in the 1970s by Richards J. Heuer, Jr., the same veteran intelligence officer at the Central Intelligence Agency (CIA) who wrote *Psychology of Intelligence Analysis* (Heuer, 1999). ACH is now used reasonably widely, not only in the intelligence community, but also in other fields in which one must make the best judgment on the basis of uncertain data.

The basic idea behind ACH, Thomason explained, is that there is a tendency among experts—as, more generally, among all of us—to ignore certain hypotheses and certain data when trying to make sense of a situation. The natural approach is to choose a hypothesis that seems most likely to be true and to see if the data support it. If supported, then the hypothesis is assumed to be correct; if not, the next most likely hypothesis is examined.

In contrast, the first step in ACH is to list at the outset all possible hypotheses, preferably by working with a number of people with different perspectives. Then one lists all the arguments and data that support or rebut each of the hypotheses. This process forces analysts to consider all of the hypotheses and all of the evidence. A series of steps follows: deciding how useful the various arguments and bits of evidence are in deciding among the hypotheses, using the evidence and arguments to attempt to disprove the various hypotheses, forming a tentative conclusion, analyzing the sensitivity of the conclusion to various key bits of evidence, and considering the consequences if particular pieces of evidence are wrong. After that, there follows a discussion of various hypotheses, not just one, and how likely each is to be correct.

Heuer's belief, Thomason said, was that this approach would force analysts to pay attention to various alternatives, including hypotheses that they might otherwise ignore or play down and data that did not fit with their preferred theories. And many intelligence analysts do indeed believe that their predictions are much better because of ACH. But there

are very few studies to back this up. "Having looked at the literature, it just is not a proven process," Thomason said. "As far as I can tell, the evidence is scant. There haven't been many tests of it."

To begin with, Thomason said, there is not a single ACH approach. Over time a number of variants on Heuer's original approach have appeared, and Heuer himself has continued to modify the approach. For example, he initially called for a group of analysts to cooperate in generating a list of hypotheses, at which point a single analyst could take over. Others have since modified ACH to have such cooperation throughout the entire process. This might work better than the original method, or it might not—the issue is an empirical one that needs to be tested. More generally, Thomason commented, since there are a large number of different ACH approaches, evaluation requires a large number of tests.

A second question is, which sorts of people will find ACH most useful? Is it good for all analysts? Under all conditions? It might be, for instance, that ACH works best for experts because they are most likely to be dogmatic and in love with their own favorite hypotheses, even though they know many alternatives. Or it could be that it works best for novices, because it pushes them to think of more alternatives than they would otherwise imagine. Or maybe it is counterproductive for novices, since they don't know enough to eliminate certain hypotheses and so their final products would be almost contentless. It is unclear.

In short, Thomason said, the question that needs to be asked is not, Does ACH work? but rather "For what situations (if any) and what types of people (if any) and under which conditions (if any) does a particular approach to ACH improve analysts' expert judgments?"

Thomason said that the bottom line of the few studies that have been carried out apparently is that the approach has some promise and some problems. "It certainly isn't obvious to me that it works. It certainly isn't obvious to me that under certain circumstances it might not be counterproductive. I don't know."

It is frustrating, he said, that although ACH was originally proposed a third of a century ago and although it has achieved "a cult-like status" in the intelligence community, there have been so few studies that have tested whether and under what circumstances it actually works to improve the predictions of analysts. In part because of the intellectual isolation of the intelligence community, Thomason added, few researchers in informal logic or other areas pay attention to ACH, and so the normal scientific process that takes place in academia when a new theory or approach is suggested has not happened with ACH. The few experiments that were performed on ACH were not followed up, and the interesting results that come out of those experiments have not been developed. To fully understand the strengths and weaknesses of the various forms of

ACH, Thomason said, research conducted by the intelligence community must be supplemented by outside academic research. "It seems to me," he said, "that limiting research on ACH and other structured techniques to the intelligence community probably means this [lack of serious science] will just go on indefinitely."

It would be straightforward to have ACH and other structured-thinking techniques tested by the general scientific community of psychologists, Thomason suggested. It is simply a matter of providing the funding. And since the basic psychology of ACH should work equally well on problems outside the intelligence community, it should be possible to perform the tests in various settings with various types of participants. This should offer insight into which settings and for which types of users ACH is most effective in improving expert judgment.

Once these tests have been done, researchers can move on to field evaluation in the intelligence community. This should be relatively straightforward, Thomason said, and could be carried out in a variety of arenas that would not necessarily need to fall inside the intelligence community. One could, for instance, repeat the study on predicting Supreme Court decisions but do it with two groups: one group of legal specialists trained in ACH and a second group without ACH training. The results would be easy to interpret: if ACH works, the group using it should make more accurate predictions.

During the discussion session, Frank Stech from the MITRE Corporation agreed with Thomason that validation is important but suggested that it is understandable that ACH has not yet been validated despite having been developed more than 30 years ago. "Validation has been difficult in a number of fields," he said, mentioning medicine as another area in which field studies are difficult to design. "We don't need to just pick on the intelligence analysts." Thomason responded that specialists in these other fields generally make a concerted effort to evaluate techniques even if the studies are difficult to design and perform. It is the lack of effort that has set the intelligence community apart, he said. Even when a study has produced interesting results, people have failed to pay attention or to follow up on it. "That is just way below scientific standards in academia."

Randolph Pherson of Pherson Associates, stating that he is publishing a book with Heuer on instructional analytic techniques, noted that the largest chapter in the book is a detailed discussion of the importance of field validation. Indeed, Pherson said that he and Heuer are preparing some proposals and recommendations for strategies for performing validations of ACH and other techniques. Heuer himself recognizes that the techniques are ultimately of little use if there is no proof that they actually work, Pherson said, and so Heuer has pushed for field evaluation of those techniques.

In response, Thomason commented on the fact that, even with Heuer's urging, the various versions of ACH have never been seriously evaluated. It is testimony to just how difficult it has been to have methods used by the intelligence community assessed in the field.

Applied Bayesian Analysis

A second approach to improving prediction applies Bayesian analysis, a statistical approach that uses observations and other evidence to regularly revise and update a hypothesis. Charles Twardy of George Mason University described APOLLO, a software application that uses an advanced form of Bayesian analysis called Bayesian network modeling to help analysts predict the likely behavior of a country's leader or other persons of interest.

There is nothing new about using Bayesian analysis to improve prediction, Twardy said. From 1967 to 1979 the CIA had at least one active research group applying basic Bayesian analysis to making predictions. The analysts used their own intuition to assign an initial probability to an event and then modified that probability, either intuitively or mathematically, as certain events happened or failed to happen. The main difference between the Bayesian approach and the intuitive approach is that, for the Bayesian approach, the analysts had answered a series of what-if questions ahead of time about the probabilities of something being true if something else happened. Then, as events unfolded, the Bayesian method modified the initial estimated probability according to a mathematical formula that depended on the answers to the series of questions. Thus, although the Bayesian analysis did depend on input from the analysts, the analysts were not directly involved in modifying the probability over time.

One of the initial tests of Bayesian methods in intelligence analysis was a retrospective one. Some years after the end of the Korean War, the CIA research group examined the events leading up to the massive Chinese invasion of North Korea on November 25, 1950. At the time, the invasion by the Chinese caught the United Nations and U.S. forces completely off guard. However, Twardy said, a retrospective analysis using the Bayesian method applied to evidence available in mid-November estimated that the odds were three to one that the Chinese were about to intervene in the war on a large scale. If such an analysis had been performed at the time, the United States would not have been surprised by the invasion.

The retrospective analysis was more suggestive than convincing, Twardy noted, since things always seem clearer in hindsight. What was needed was evidence that it is possible to come to the right conclusion ahead of time.

Such evidence appeared with an experiment involving five analysts predicting the likelihood in mid-1969 that the Soviet Union would attempt to destroy China's nascent nuclear capabilities. The five analysts used both conventional and Bayesian methods to estimate this likelihood over a period stretching from late August to late September 1969. In the conventional approach, the analysts assigned numerical probabilities based on their own judgment and intuition. The five analysts started out with varying estimates of the probability of a Soviet action—anywhere from 10 to 80 percent—and, as events unfolded throughout September 1969, all of them revised their estimates steadily downward until their estimated probabilities were close to zero.

For four of the five analysts, Twardy said, applying Bayesian statistics led them to the same conclusion—that the Soviets were not going to go to war with the Chinese—but it got them to that conclusion much more quickly. The results for two of the analysts are shown in Figure 2-1. "The Bayesian method generally made analysts revise their estimates faster than they would have intuitively done."

Strangely, the Bayesian analysis performed by the fifth analyst (shown in Figure 2-2) had vastly different results from the analyses of the other four and by the end of September was still predicting a 75 percent chance

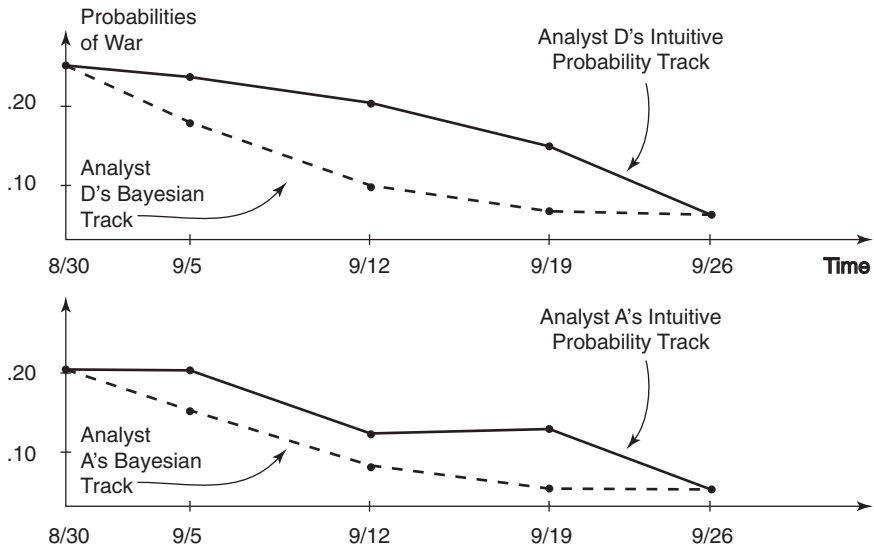

FIGURE 2-1 Probability tracks for Analysts A and D using both conventional and Bayesian methods.
SOURCE: Fisk (1972). Reprinted with permission.

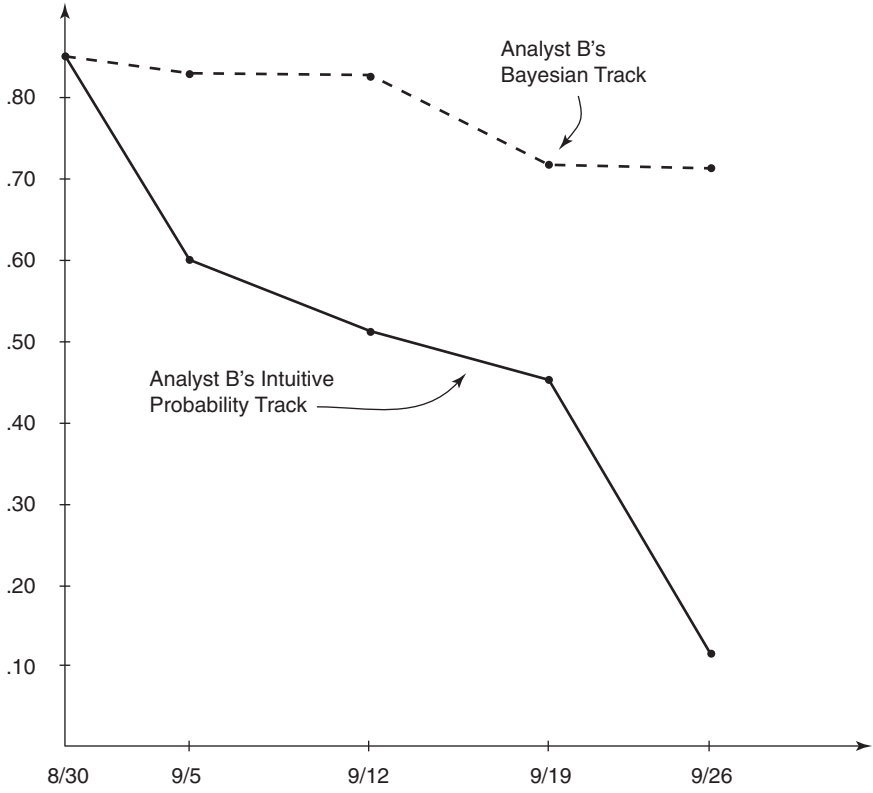

FIGURE 2-2 Probability tracks for Analyst B using both conventional and Bayesian methods.
SOURCE: Fisk (1972). Reprinted with permission.

that the Soviets would go to war, even as that analyst's intuitive estimate had dropped almost to zero. "Here is why you need to do experiments," Twardy commented. "Even though there are tremendous numbers of studies in the psychology literature that you are better off with Bayes, you never know what is going to happen when you give it to analysts." It seems likely that the fifth analyst had somehow misunderstood how to apply the Bayesian analysis—and that is an important piece of information that only a field study could detect.

The research group's work ended in 1979, and Twardy said he has seen no evidence that work with Bayesian methods continued beyond that. To all appearances, Bayesian analysis essentially vanished from intelligence analysis.

Recently, with the development of APOLLO, a descendant of those

original Bayesian analyses has appeared. APOLLO, created by Paul Sticha and Dennis Buede of HumRRO, and Richard Rees of the CIA relies on Bayesian networks instead of the much simpler Bayesian statistical calculations used by the CIA research group in the 1970s, but the goal is the same: to help analysts overcome their biases and improve the accuracy of their predictions by changing their probability estimates in response to things that happen or do not happen over time.

The standard example of how APOLLO can be put to use is in predicting what the leader of a foreign country is going to do in response to overwhelming labor strikes. Will he leave the country? Will he respond with violent repression of the strikes? To answer that question, a highly detailed "decision model" (see Figure 2-3) is created, ideally with the input of a number of analysts and experts assembled for a two-day meeting. The model traces how various eventualities affect one another, with probabilities assigned for each cause-and-effect relationship.

Once the model has been defined, all an analyst needs to do is to

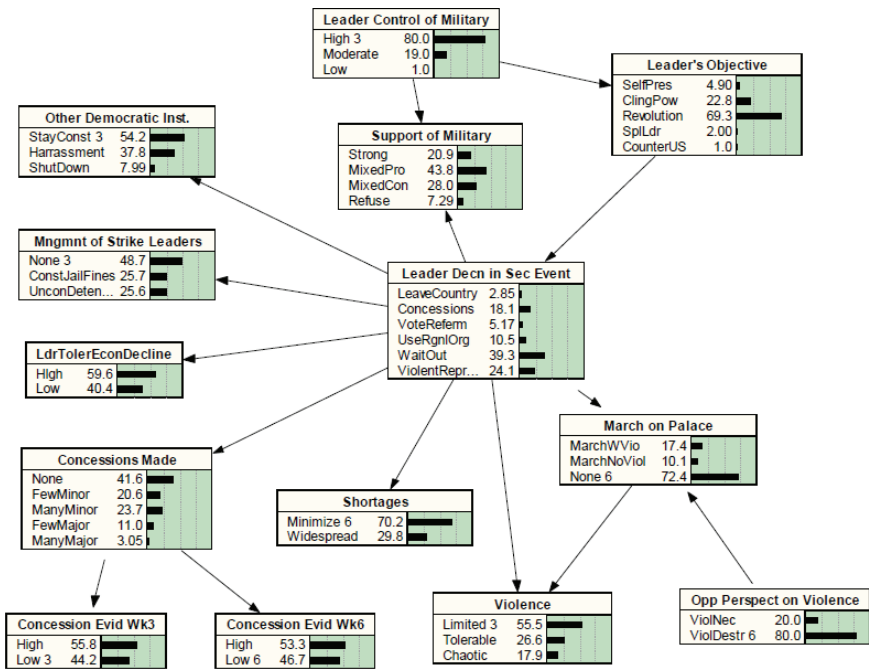

FIGURE 2-3 Example of a situation model. The hypothesis node is in the middle. SOURCE: Sticha, Buede, and Rees (2005). Reprinted with permission.

monitor the developing situation and input which events have happened. Perhaps six weeks after the strikes, the leader has responded with a few concessions, and there has been limited violence. The analyst checks off those possibilities in the model, does nothing with the possibilities that are still unknown, and the software automatically revises the probabilities of what the leader will do.

The model also offers other results of the analysis. It is possible, for example, to examine which events the predictions depend on most strongly and which events make little difference to the predictions. The model can also be used to work backward from observed actions to the motivations of the leader. For example, if a leader's major goal is to stay in power, his responses to various events should be significantly different than if his goal is, say, to move his country toward democracy. The model helps untangle the various causes and effects and focus on what events imply about motivations.

More generally, APOLLO includes a psychology module that fits into the larger model and takes into consideration the self-image and other psychological characteristics of the leader. Psychologists can assess the various personality factors of a leader—by examining the content of speeches, for example—and then input these factors into the model to improve its predictive power.

APOLLO could be valuable to analysts in a number of ways, Twardy said. Its main goal is to improve predictive power, but it could be put to work in other ways as well. It could be used, for instance, to determine which pieces of information would be most useful in improving a prediction, so that intelligence collectors would know where to direct their efforts for the biggest payoff. Or, as noted above, it could help indicate what a leader's motives are.

The group that developed APOLLO is now testing it, Twardy said. The original idea was to perform a randomized controlled trial, but they abandoned that plan for a couple of reasons. First, they did not think it would be possible to sign up as many analysts as they would need to get statistically significant results. Second, in a controlled trial, half of the analysts would not be assigned to use the method they had signed up for—APOLLO—so many of them would likely walk away before the end of the experiment, resulting in a bunch of annoyed analysts and statistically marginal results at best.

So the APOLLO group settled on a pre- and post-test design in which the probabilities are tested both before using the method and after, and one looks to see if using the method has improved the outcome. This is a workable approach, Twardy said, but it will take a long time to gather the data. Furthermore, the test would be most useful if the gain from the

APOLLO method was compared with gains from other methods, such as ACH, but that will require more work and more analysts.

One of the obstacles to evaluating analysts' use of APOLLO—or even to convincing analysts to use a structured method like APOLLO in the first place—is how busy most of them are. Their days are filled by the responsibilities they already have, and unless something is done to give them more time or to restructure their incentives, Twardy said, they are unlikely to adopt something like APOLLO, much less participate in an evaluation of it. "So it might not even be feasible to do real field evaluations with analysts who are so pressed for time," Twardy said.

As an alternative, he suggested, it might be possible to test APOLLO or similar techniques in the intelligence academies. The tests would not be as realistic as if they were carried out with working analysts, but at least the subjects would be analysts in training, and it would be much easier to get access to them. If that doesn't work, tests could be carried out in professional schools, such as business schools, where they do forecasting. Or the tests could even be carried out with psychology undergraduates, the usual lab rat of psychology experiments. "You sacrifice more validity, but you get a lot more access." And, he continued, there is no reason it wouldn't be possible to do a tiered approach, in which much of the early work is done with psychology undergraduates, and then, after the bugs are fixed, further testing is done in the intelligence academies or with working analysts.

Finally, Twardy noted, one of the obstacles to incorporating APOLLO or something similar in the work of intelligence analysts—in addition to the extra time it demands from them—is the fact that it requires quantitative estimates of probabilities: there is a 20 percent chance that the striking workers will back down if the army is called in; there is a 40 percent chance that the army will take over the country if the leader flees; and so on. But although there has been some recent progress, Twardy said, it has traditionally been very difficult to get analysts to assign numerical probabilities to their estimates. They are more comfortable with words like *probable*, *unlikely*, or *near certain*. Unless they move from the qualitative to the quantitative in their predictions, APOLLO and similar techniques will be out of reach.

3

Field Evaluation Experiences in Other Areas

Field evaluations are performed to one degree or another in many other areas besides intelligence and counterintelligence, sometimes effectively and sometimes not. Five speakers in two separate sessions discussed their experiences with field evaluation in other areas with an eye toward offering takeaway lessons that could be applied to the field of intelligence.

EDUCATION

The field of education has many parallels with intelligence: its practitioners have a large body of accepted methods, many of which have not been rigorously field tested; new methods are regularly introduced, many of them based on behavioral science research, but, again, a large percentage are never carefully evaluated in the field; and the practitioners tend to trust their own experience-based judgment over research findings that may be contradictory. Grover Whitehurst, who served as the first director of the Institute of Education Sciences in the U.S. Department of Education (ED), described the experiences of ED during his tenure[1] in attempting to subject educational programs to rigorous evaluation, emphasizing randomized controlled trials (RCTs).

[1]The extent to which randomized controlled trials should direct education policy and practice has been the source of considerable debate, particularly if it excludes other research methods. Recently, ED began a reanalysis of its research and evaluation priorities.

"There are always barriers to generalizing from one field to another or from one agency to another," he said. "I am acutely aware of that. But I think there may be some lessons learned in other federal agencies about the production and use of research that could be applicable to intelligence."

Whitehurst, now a senior fellow of government studies and director of the Brown Center on Education Policy at the Brookings Institution in Washington, began his presentation by commenting that education has a history of not knowing much about what it is doing. In 1971, for example, a RAND Corporation group set out to discover which educational practices were well supported by research and came to the conclusion that there was essentially no evidence to show that any educational methods worked. Almost three decades later, a report by the National Research Council reached a very similar conclusion (National Research Council, 1999). So there had been very little progress in 30 years.

A 2007 review of all federal investments in science or mathematics education listed 105 programs with statutory funding authority of $3 billion a year (U.S. Department of Education, 2007). According to Whitehurst, when the Office of Management and Budget instructed each of those programs to submit any evaluations they had of the effectiveness of their investments, the only positive results that anyone could find came from four small research grants, such as one in which an education researcher at Carnegie Mellon University had evaluated a technique for teaching students the principles underlying experiments and found it to be effective.

The lack of validation was not limited to the education programs in science and mathematics, Whitehurst said. It was an issue throughout the federal government. For instance, when he attended an Education and Training Conference at the Department of Labor, in preparation for the meeting, he searched through reports from the Government Accountability Office for information on education and training programs that the Department of Labor had sponsored. He could not find a single evaluation of any recent program that contained rigorous evidence that the program had had an effect.

One of the main reasons that so few RCTs of federal programs are carried out, Whitehurst said, is that there are few people in the agencies who understand the importance of evaluating programs. He described serving on a committee organized by the White House science office to study what needed to be done to improve the evaluation of science, technology, engineering, and mathematics programs. As part of that process the committee members met with representatives of various federal agencies involved in funding such programs and asked them to describe the programs they wished to expand and to discuss their evaluation plans for

those programs. What became clear to the committee members, he said, was just how little understanding there is in many federal agencies of the importance of asking whether an investment is actually working.

"So there is a strong sense," he said, "that for federal agencies to do a better job in producing knowledge about the effectiveness of their own program activities, they need people embedded in the agency who understand the importance of doing that, who have the credentials and a position in the agency that enable their aesthetic to have some impact."

Over the past decade, there was a move in this direction at ED, and, as a result, an increasing number of effective evaluations were performed on ED-sponsored programs. In 2000, Whitehurst said, there were 17 evaluations of ED programs, but only one of these evaluations had a randomized design that made it possible to rigorously answer questions about the effect of the program. Instead, an evaluation was more likely to be based on such things as surveys sent to principals asking if they thought the funding they received had been helpful. Not surprisingly, such evaluations generally found evidence of a positive effect.

In more recent years, ED has moved toward the use of randomized trials to measure the effect of programs or, if a randomized trial was not feasible, using the best available approach. There is now an emphasis on comparative effectiveness trials, measuring the effects of different approaches against each other. As an example, Whitehurst described an evaluation of four elementary school mathematics curricula; the research found that the difference between the most effective and the least effective of those curricula was equivalent to almost a half-year of learning over the course of a school year. And since the costs of the different programs were about equal, that is a valuable finding—one that would not have been possible to reach without a well-designed comparative effectiveness evaluation.

Today, Whitehurst said, ED has a list of 70 programs and practices that have a strong evidentiary base behind them and that have been shown to have a positive effect on student outcomes.

How did ED turn things around and start to evaluate programs rigorously and regularly? There were a number of factors, Whitehurst said. First, in 2002-2003 Congress created the Institute of Education Sciences. Because the institute was given a degree of independence that its predecessor agency did not have, it was able to attract knowledgeable people to perform good research and protect that research from outside pressures. There was also more money provided for randomized controlled trials.

The experience at ED offers a number of lessons for the intelligence community or for any group that wishes to begin evaluating methods and techniques in a rigorous way, Whitehurst said.

One of the things that makes a difference is money. Without a predict-

able supply of funding, researchers will not get involved, nor will they train graduate students in the area.

The methods used in the research also matter, he said. "If we are going to try to answer a question that policy makers are supposed to run with, it is important to have an answer that they can stand on, and not one that in many cases will be incorrect." One of the things that ED has studied, for instance, is how to arrive at the right answer with a variety of methods. The more rigorous the method, the more quickly one can come to a conclusion that can be depended on. Other methods require replication and aggregation to overcome the potential errors and omissions and arrive at a reliable answer.

One problem that ED faced was that the quality of the research on education varied tremendously, Whitehurst said. Part of the solution was the establishment of the What Works Clearinghouse by the ED's Institute of Education Sciences. The clearinghouse assesses the quality of research and confers its imprimatur on those studies it sees as being sufficiently rigorous.

Independence matters. In government there is generally strong pressure to support policy. Policy makers decide on policy, and they do not appreciate evidence that their favored policies do not work. Thus an agency that is producing knowledge needs to have a certain amount of independence to be able to resist the pressures to come up with conclusions that support policy.

Finally, he said, people matter. It is important to have people involved who understand science and who are committed to the value of the scientific effort and its integrity. One approach to doing this is to get the academic research community involved, and the way to do this is to provide stable funding with a reasonable peer review process.

CRIMINAL JUSTICE

In her presentation, Cynthia Lum, deputy director of the Center for Evidence-Based Crime Policy (CEBCP) at George Mason University, discussed the current state of field evaluations in two areas of criminal justice: policing and counterterrorism. In neither area is the field evaluation of methods and technologies yet standard practice, although such evaluation is much closer to reality in one of the areas than in the other.

In most police departments today, Lum said, decisions about policing practices and policies are not based on scientific evidence but instead are made on a case-by-case basis, using personal experience, anecdotes, even bar stories as a guide. The individuals making the decisions fall back on their own judgment, guesses, hunches, feelings, and whims, while being influenced by various outside forces, such as political pressure, lobbying

by special interest groups, social crises, and moral panics. "That is the reality," Lum said. "The context of decision making is not crime, it is police organizational culture [and] things that have nothing to do with crime."

That is slowly changing, however. There is a growing emphasis on what is called "evidence-based policing," which is the practice of basing police policies and procedures on scientific evidence about what works best. Many of the ideas for evidence-based policing were laid out in a 1998 article by Lawrence Sherman (Sherman, 1998), now the director of the Jerry Lee Centre of Experimental Criminology and the Police Executive Programme at Cambridge University in England. Sherman talked about using various data accumulated by police departments, such as maps of crime patterns and information about repeat offenders, combined with scientific methods of deduction and objective evaluations of programs to shape police policies and methods. The goal was to replace subjective judgments with scientific conclusions about what works best in determining what police practices should be.

Over the past decade there has been a slowly growing acceptance in police departments of the importance of this scientific approach to policing, Lum said. The aspect of evidence-based policing of greatest relevance to the workshop is, of course, the scientific evaluation of different policing practices. Such evaluation is beginning to be done, Lum said, but its pace of adoption varies greatly according to what is being evaluated.

For example, very few of the new technologies being used in policing—things like license plate recognition technology, crime mapping, mobile computer terminals, and DNA field testing—have been evaluated for effectiveness. They are tested to see if they work as they are supposed to—to determine, for instance, if the license plate recognition systems actually do recognize license plates accurately—but almost nothing has been done anywhere to test if these technologies are effective in, say, reducing crime rates. "I actually don't know of any very high-quality evaluations in police technology," Lum said. Evaluations currently being completed by the Police Executive Research Forum and by CEBCP are looking at whether license plate recognition systems have any effect on reducing auto theft—which is what they are intended to do—but these are the first studies of their kind, so far as Lum knows. Most technology "evaluations" have examined whether the technology physically works or is faster, but not necessarily more effective, she said.

However, there have been quite a few evaluations of various policing interventions, such as hot spot policing, crackdowns, raids on crack houses, and other actions designed to reduce the frequency and severity of crimes. Lum and her colleagues have assembled, in an "Evidence-

Based Policing Matrix,"[2] a set of 92 evaluations of policing interventions that were judged to be medium to high quality. Many of the studies were randomized controlled trials or "high-level quasi-experiments," Lum said, and they looked at a variety of outcomes, such as crime rates, recidivism, even police legitimacy.

More generally, she said, there is reason to be optimistic that there will be more and more high-quality evaluations of policing practices in the future. The reason for such optimism is that a research infrastructure already exists in the area of policing that can be used for the generation of scientific evidence and objective evaluations.

This research infrastructure exists in large part, she said, because of the efforts of a number of pioneering researchers who spent their careers working to build relationships and trust with police officers and departments around the country. Because of their efforts, there are police chiefs and officers in many police departments who recognize the value of evidence-based policing and of performing evaluations and who therefore are willing and able to cooperate with policing researchers.

A second important part of the research infrastructure is the knowledge base, exemplified by the 92 evaluations of interventions that Lum and colleagues collected. Such a knowledge base is vital in determining what questions to ask, what areas to focus on in developing better evidence, and where to spend research funds.

The research infrastructure also includes advances in information technology that are being applied to policing. Lum mentioned, for instance, that she works with a geographic information system that allows her to map the locations of tens of thousands of crimes very quickly and easily.

Finally, shifts in police culture have been vital in developing and using this research infrastructure. Not only are an increasing number of police chiefs and officers recognizing the value of science and research to their jobs, but police research groups, such as the Police Foundation and the Police Executive Research Forum, are receptive to working with academic criminologists. Those in the policing community are more understanding of the value of embedding a criminologist in a police department, for example, and the professional groups are now working to translate the criminological research for their members and to explain its importance.

In contrast to the situation with regard to policing, Lum said, the research infrastructure and the evidence base for counterterrorism work are almost nonexistent. "It is very weak and very small."

Since the September 11, 2001, terrorist attacks there has been a tremendous investment into the counterterrorism area. Funding has been shifted

[2]See http://gemini.gmu.edu/cebcp/matrix.html.

into the area, much of it from policing and emergency management, along with new funding, and there has been a great deal of discussion about tools and technologies to prevent and counter terrorism as well as new policies and laws intended to make terrorist attacks less likely to succeed. At the same time, there has been a surge in publications about various counterterrorism subjects, such as airport screening and metal detectors, detection devices for biological or chemical weapons, emergency response preparedness, hostage negotiation, and many others.

In 2006 Lum and two colleagues published a systematic review of the literature on counterterrorism (Lum, Kennedy, and Sherley, 2006). They surveyed more than 20,000 articles written about terrorism and counterterrorism. Of all of those articles, Lum said, only seven contained evaluations that satisfied minimal requirements for methodological quality. In other words, although there has been a torrent of literature in the area, much of it discussing new technologies and practices for dealing with terrorism, almost none of it offers useful evaluations of these technologies and practices.

Even more disturbing, Lum said, is the fact that there is almost no research infrastructure in the counterterrorism area to support an evaluation agenda. Many of the studies are done by the same people using the same datasets that are just added to over the years, allowing the researchers to analyze and reanalyze them from varying perspectives. Thus, despite the large number of publications, the evidence base is inadequate.

There are also few research pioneers in this area, Lum said, and most of them tend to focus on such things as the causes of terrorism, the psychology of terrorism, and the groups behind terrorism. There is far more work done in such areas than in addressing the question of whether various interventions work.

One reason that the area of counterterrorism, unlike policing, has so few evaluation studies is the fact that terrorist activities, unlike criminal activities, are rare. Thus researchers have relatively little information to work with. Researchers also sometimes find it difficult to obtain the clearance they need to gain access to information, and relatively few research relationships have been developed between practitioners and researchers.

Furthermore, Lum said, she has found that the leadership culture in counterterrorism is not focused on science or on the role that science could play in counterterrorism. Nor are there any third parties that can play the kind of role that the Police Foundation or the Police Executive Research Forum plays in advancing the evaluation of policing practices.

Finally, there is little or no government support for evaluation research. Much of government funding is instead focused on the causes of terror-

ism and descriptions of relevant technologies rather than information about interventions. This is exacerbated by a lack of discourse or rhetoric about evaluation or science as they apply to counterterrorism methods.

HEALTH SCIENCES

In the next presentation, Lisa Colpe, a senior scientist and epidemiologist at the National Institute of Mental Health, described a pair of mental health surveys that illustrate the importance of including up-front evaluation criteria in research studies.

One of the ways to study the epidemiology of a mental illness in the general population is to conduct surveys of people at home or in other specific settings. But the people responding to such surveys will usually give researchers only a limited amount of time to ask their questions, so the surveys must be designed to detect the signs of illness with a set of questions that is much shorter than the standard clinical questionnaire. Depression, for example, is normally diagnosed in a doctor's office after a lengthy examination and interview, but researchers who are interested in determining the prevalence of depression in the general population must develop an abbreviated questionnaire that can predict with a certain accuracy which of the respondents has depression.

Colpe described a recent survey that she and a group of colleagues conducted to look for serious mental illness in the general population. By definition, a serious mental illness is characterized by the presence of a mental disorder combined with significant functional impairment. The researchers developed two scales for use in the survey, one that indicated the presence of mental distress and the other that measured the degree of functional impairment. By combining sets of responses from the two scales, they could acquire a measure of the prevalence of severe mental disorders in the study population.

It was then necessary to compare the measure of mental disorder derived from the survey responses with the measure of mental distress derived from standard clinical interviews. In particular, psychologists use the Structured Clinical Interview for DSM-IV disorders, or SCID, to detect the presence of a clinically significant mental disorder; the researchers needed to calibrate their survey responses with the SCID responses so that they could be sure they were compatible.

The process was straightforward. The researchers took a sample of 750 people who had answered the survey questions and assessed them clinically with both the SCID and the Global Assessment of Functioning Scale, which is a standard measure of impairment. With these measures they could assess which of the 750 subjects had a serious mental illness

according to clinical standards and then examine how well these particular subjects were identified by the set of questions in the questionnaire.

In particular, Colpe said, they found that not every item on the questionnaire had the same predictive weight in picking out those subjects with a serious mental illness according to clinical criteria. A statistical analysis allowed them to assign different weights to the survey questions to get the most predictive power from the questionnaire. "What we were looking for," she said, "was a cut point where we were happy with the estimate that we were getting by evening out false positives and false negatives and coming up with an estimate of serious mental illness that could be applied to the greater survey." By applying those numbers to all 45,000 adults who took part in the survey, they were able to get estimates of the prevalence of serious mental illness throughout the United States because the survey sample was chosen to be nationally representative.

This validation will be carried out continuously, Colpe said, because the survey itself is an annual one. The researchers carrying out the survey will routinely select a small subgroup of survey respondents to be given a full clinical evaluation so that the validation remains current.

It is important to build this sort of validation or calibration into larger surveys, Colpe said. Even when researchers believe they understand their measures, it is possible for those measures to vary in unexpected ways. For example, the answers that people give to survey questions will often vary depending on whether the questions and answers are given orally or in written form. "People respond differently to sensitive items about their mental health if they are being asked by a grandmotherly type sitting across the table versus if they are able to answer the questions on a computer or in some way that is a little more discreet."

Colpe also described an international collaboration, the World Mental Health Survey Initiative, carried out by researchers in 28 countries located in all the different regions of the world and surveying a total of more than 200,000 people. The goal of the survey was to estimate the prevalence of various mental disorders, the accompanying societal burdens, the rates of unmet needs, and the treatment adequacy in the different nations. One of the major challenges was making sure that the results would be comparable from country to country.

It began with the design of the study. All of the participating countries agreed to one universal design in which there would be nationally or regionally representative household surveys. The various countries also agreed to use the same training and quality control protocols as well as the same processes to translate the survey instrument into the different languages. Finally, all of the countries agreed to do clinical validation studies of the sort described above that would be used to validate the responses on the surveys. As a result, Colpe said, the differences that the

survey uncovered among countries were likely to be reflective of real differences rather than being artifacts created by differences in how the surveys were carried out in the various countries.

"The bottom line message," Colpe said, "is that you have got to plan on the evaluation from the beginning of the study as you design the study, so you are well positioned to do comparisons across studies or conduct pre- and postintervention program analysis."

LEGAL SYSTEM

Over the past two decades there has been a great deal of testing and evaluation of one particular aspect of the U.S. legal system—the use of eyewitness identifications. Christian Meissner, an associate professor of psychology and criminal justice at the University of Texas at El Paso, provided an introduction to the general field of psychology and law, which includes the specific topic of eyewitness recall identification, and discussed how research has helped lead to reforms of the legal system.

The area of psychology and law, Meissner explained, includes experimental psychologists with social, cognitive, developmental, and clinical backgrounds who conduct basic and applied research with the goal of helping improve the legal system. The field has evolved and expanded over the past 100 years. One of the earliest practitioners was Hugo Munsterberg, often considered the father of applied psychology. In 1908 Munsterberg published *On the Witness Stand*, which explored some of the psychological factors that could affect the outcome of trials, such as the variability of eyewitness testimony and the phenomenon of false confessions (Munsterberg, 1908). Now considered a landmark, at the time the book was controversial and soured the relationship between experimental psychologists and the legal profession for many years, Meissner said.

So despite Munsterberg's pioneering contributions, there was little sustained research in the field for nearly 50 years. Part of the reason, Meissner said, was simply that psychologists did not have the methods or theories that allowed them to contribute to the legal system. It was not until the 1950s, for instance, that cognitive and social psychology began to be developed. By the 1960s psychologists were starting to apply cognitive and social theories to the real world and, in particular, to the legal system, but it has only been in the past two or three decades that this work has kicked into high gear.

The trigger for much of this work, Meissner said, was the first case of DNA exoneration, which occurred in 1989. Indeed, there were three separate cases in 1989 in which DNA evidence was used to prove the innocence of someone who had been convicted by the criminal justice system. "This was a really important moment in our field," Meissner

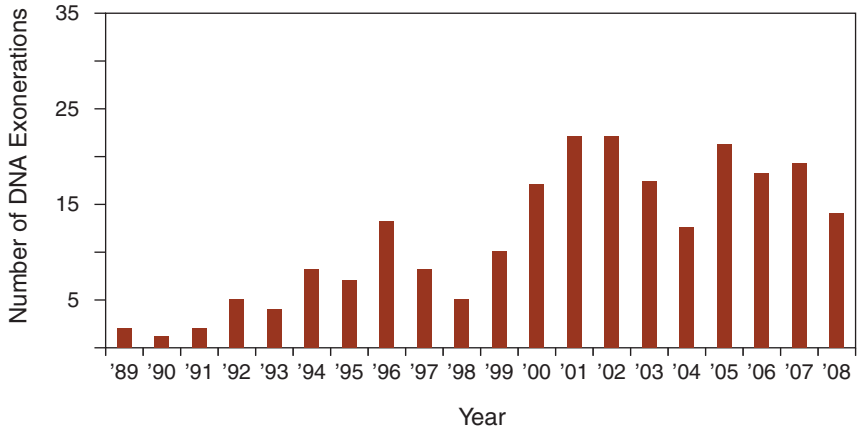

FIGURE 3-1 DNA exonerations in the United States.
NOTE: The Innocence Project regularly updates these numbers, which have grown even more since the workshop. For the most recent figures, see http://www.innocenceproject.org/know.
SOURCE: Innocence Project (2009a). Reprinted with permission.

said, "because it provided impetus not only for further research but also for reform." Since that time, according to data from the Innocence Project, the total number of people proved to have been wrongfully convicted through DNA exoneration has grown to more than 240 (see Figure 3-1), including 17 on death row.

A variety of studies were carried out to determine the various causes of wrongful conviction, Meissner said, and they have found that mistaken eyewitness identification played a role in about 80 percent of the cases (see Figure 3-2). That shocked the legal system, he said, and led to a great deal of research into witness identification and efforts on validating its use.

The research on eyewitness memory has actually been going on in earnest since the late 1960s and early 1970s, and today the literature includes hundreds of studies on interviewing witnesses and eyewitness identification. These studies include such research as laboratory studies on face recognition and more formal studies of eyewitness performance, such experiments in which a witness observes an event and later is asked to report what happened.

In investigating eyewitness reports, Meissner said, a key distinction is between system variables and estimator variables. System variables are those that are controlled by the system. They include such things as the ways that interviews are conducted, the ways that identification lineups

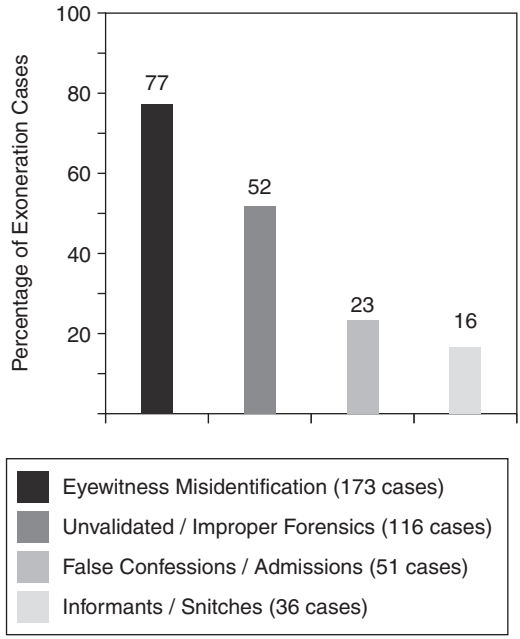

FIGURE 3-2 Contributing causes of wrongful convictions in the first 225 DNA exonerations.
NOTE: Total is more than 100 percent because wrongful convictions can have more than one cause.
SOURCE: Innocence Project (2009b). Reprinted with permission.

are created and administered, and the types of instructions that witnesses are provided. Research on these variables is important because it can help identify ways that the system can be modified to reduce the inaccuracies of witnesses and provide more reliable results.

Estimator variables, by contrast, are those over which the system has no control but that can still affect witness memory. Meissner offered several examples: How good a view did the witness get? How long did the witness have to study the person's face? What was the length of time between the witness viewing the suspect and the witness providing an identification? It is important to understand these variables as well, Meissner said, not to change the system but to aid in assessing the reliability of a witness.

One of the key facts about the research on eyewitness performance is the fact that it has taken a multimethodological approach, Meissner said. It has included well-controlled experimental studies done in laboratories in which key variables were manipulated and the outcome observed. It

has included a variety of field experiments, such as a situation in which the participants actually went into a convenience store and experienced something and were asked about it later or in which the participants were put in the role of a clerk at a convenience store or a bank who is passed phony bills and then later asked to identify the person who passed the counterfeit money. It is possible to control a number of the variables in these situations, so they can be considered quasi-experimental. There have also been archival studies, in which researchers looked back at real cases and attempted to determine the factors that influenced whether a person was selected as a suspect. More recently there have been evaluations in which researchers looked at how changes in policy affect indicators that are important for the legal community.

Ultimately, Meissner said, this research has led to the development of a number of procedures to improve eyewitness identification. For example, interviewing protocols have been developed that dramatically increase the amount of correct information that witnesses recall without causing concomitant increases in errors. Researchers have also proposed a number of improvements in identification procedures, such as how best to construct a lineup, the use of double-blind administration in carrying out lineup identifications, the best way to give instructions to witnesses, and the use of confidence assessments in determining how sure witnesses are of their identifications.

An important fact to note about this body of research, Meissner said, is the sheer amount of it. There have been a number of meta-analytic reviews of the studies. Indeed, there has been so much research in some areas that two or three meta-analyses have been done.

The research is not without some controversy, however. Meissner described a technique, the sequential lineup, developed in 1985 by Rod Lindsay and Gary Wells (Lindsay and Wells, 1985). The idea behind the sequential lineup is that instead of showing a witness a group of six people or photographs all at one time, the witness would be presented with the people or photos one at a time, and at each point the witness has to determine whether or not this is the person who had been seen. The first couple of studies done by Lindsay and Wells found a dramatic drop in misidentifications but no significant drop in correct identifications. The technique looked to be a major improvement.

However, subsequent studies found that the technique produced not only a drop in false identifications but also a drop in correct identifications. The technique was not making witnesses more discriminating; it was making them more conservative in their identifications. They were less likely to identify any given individual, either the guilty person or an innocent one. The findings have led to controversy about whether sequential lineups should be used. They would probably lead to fewer innocent

people being found guilty, but they might also reduce the likelihood that guilty parties would be identified.

The larger point, Meissner said, is that if additional research had not been done after the early papers were published, the sequential lineup technique might well have been implemented widely, and it would only have been much later, after subsequent research was done, that it became clear what the costs of the technique really were.

In 1996, the American Psychology–Law Society commissioned a white paper by a group of leading eyewitness researchers, which made recommendations for changing the system (Wells et al., 1998). At around the same time, the National Institute of Justice (NIJ) produced a report on the first 28 cases of DNA exoneration, 25 of which involved mistaken eyewitness identification (Connors et al., 1996). Those two documents in turn led to the formation of an NIJ working group composed of researchers, prosecutors, defense attorneys, law enforcement investigators, and others in the criminal justice system who worked to provide a set of best practices for collecting evidence from eyewitnesses. The document produced by that group, *Eyewitness Evidence: A Guide for Law Enforcement*, was published in 1999 (Technical Working Group for Eyewitness Evidence, 1999). Four years later, the NIJ published an accompanying training manual, *Eyewitness Evidence: A Trainer's Manual for Law Enforcement* (Technical Working Group for Eyewitness Evidence, 2003).

One of the morals of this story, Meissner said, is that it is important for researchers to follow through to the end of the process—the actual training of the people who will put recommended practices into effect. "Sometimes when you develop procedures in the lab or even in the field," he said, "when they get implemented they get changed." Furthermore, they can get changed in ways that compromise the validity of the technique. So this is an important consideration for anyone discussing field evaluation: not only is it important to do the research and see that the findings are implemented, but also the findings must be implemented in the way they were intended.

However, Meissner concluded, the reports are just recommendations; although they serve as an example of best practices in the field, they do not have the force of law. And today the majority of law enforcement jurisdictions still have not changed their procedures. Thus, although the work on eyewitness identification has been a success in terms of the research, the evaluation, and the consensus recommendations, widespread success in implementation remains unseen.

HUMAN FACTORS

In his presentation, Eduardo Salas, a professor at the Institute for Simulation and Training at the University of Central Florida, described the field of human factors engineering and how various technologies for enhancing human performance are evaluated in the field.

Human factors is the study of humans and their capabilities and limitations, Salas said. The understandings developed from the science are applied to the design and deployment of different systems and tasks. Every device that a person uses, from an automobile to an iPod, has been shaped by human factors engineers, he said.

A closely related field is organizational psychology, which is the study of people and groups at work. Its basic application is to make organizations more effective. "This is a science that basically resides in human resources departments," Salas said, adding that anyone who has been interviewed by a human resources specialist or who has taken an assessment as part of applying for a job has been touched by an organizational psychologist.

Human factors scientists and organizational psychologists have a long history of working in fields in which errors can have drastic consequences: the military, hospitals, and the aerospace and nuclear industries. Because the stakes are so high, it pays to think carefully about how the design of devices and of organizations can make mistakes less likely.

Historically, Salas said, the military has driven much of the development in human factors and organizational psychology. During both world wars, technologies were introduced without being empirically validated for use with humans, and in many cases it was discovered too late that the technologies did not work particularly well. There was also a growing need for systems that would be usable no matter who was operating them. Thus the military began to invest in research on human factors and organizational psychology, and today the military funds much of the basic and applied research in the field.

Besides the military, medical communities today are investing a great deal in the field. There is particular interest in encouraging teamwork. At the same time, everyone wants to be sure that the various approaches being adopted actually work, Salas said. "Every CFO [chief financial officer] in every hospital is being asked the question: before I roll this out, tell me if this works. What is the evidence that you have? If you spend eight million dollars in peak training for 14,000 employees, will it work, and will patients be safe?" Evaluations are thus a vital part of the field.

Human factors and organizational psychology offer a large variety of products, Salas said. These include tools and devices, principles, strategies, methods, guidelines, and theories, and all of them are evaluated to some degree. Clients naturally want to know what the evidence is that

something will work, he said, and that pushes the human factors and organizational psychology community to do a great deal of evaluation and evidence-based reporting.

There are two primary approaches to evaluating the products, Salas said: usability analysis and training evaluation typology. Usability analysis uses human factors principles to evaluate a system and tailor it to a user, and training evaluation typology uses principles from organizational psychology and human factors to ensure that a user is able to actually use the system or perform the task and has acquired the necessary competencies.

Usability Analysis

Salas then spent the next several minutes of his presentation describing usability analysis in greater detail. "It is an iterative process that engages the user little by little in the design and the development and implementation of whatever product we are coming up with," he said. Its focus is on the system and how the user interacts with it, with the goal of improving the usefulness and effectiveness of the product.

One usability analysis, for example, focused on a problem with a particular model of car: it kept rear-ending other automobiles. The analysts came up with a variety of possible reasons, such as visibility issues, control issues, or even bad brakes. Then they tested the vehicle with different users and different tasks. What they discovered was that the gas pedal and brake were too close and kept getting pushed at the same time.

A more familiar example, Salas said, is the existence of the third brake light that has become standard on cars. "That came from a usability study done by human factors psychologists." It was found to be more noticeable to people following in other cars, thus reducing the chance of an accident when a car has to stop suddenly.

A variety of usability analyses could have application to the intelligence field, Salas said. They include evaluations of data visualization, mobile devices, wearable systems, informatics, automated decision aids, and virtual environments for military intelligence.

Five methodologies are used in usability analysis, ranging from very simple and straightforward, with the user not necessarily involved, all the way to field testing. Heuristic evaluation is the simplest and most informal. In it, usability experts judge whether a particular system fits established heuristics, like speak the user's language, be consistent, minimize memory load, be flexible and efficient, and provide progressive levels of detail. This approach is quick, low in cost, and often very effective, but users are generally not involved, so there is no user insight.

A second methodology is the cognitive walkthrough. It involves an

expert or group of experts taking on the role of the user and stepping methodically through the process of using the system. It can identify the goals, problems, and actions of the user, Salas said, and it is best used early in the usability life cycle with rapid prototypes. It has the benefit that it provides a virtual view of the user without actually having to involve one; its weakness is also that it doesn't actually involve users, because experts may miss some things that an inexperienced user unfamiliar with the system would pick up on.

The third methodology is the use of interviews with both end users and experts. This can be done in the form of focus groups, and its goal is to determine user needs and goals. If the product is already in use, the interviews can be used to determine common problems and issues from the user's perspective. The benefit of this approach is that there is direct user contact; the disadvantages are that it can be logistically complicated, and it might not be appropriate for a particular community.

The fourth methodology is the thinking-aloud protocol. This is similar to the cognitive walkthrough, except that it involves an actual user doing an actual task. The user is asked to narrate his or her thoughts while performing the task: "I need to copy this file, so I am going to look for the clipboard function. . . . Maybe it's under 'Edit' like on Windows. . . . Oh, here's a button." The purpose is to identify the user's goals, actions, and problems during actual use. "You would be surprised how much information you get out of this," Salas said. "We have a lot of evidence to show that experts cannot articulate what they know very clearly" because they know what they are doing so well that it becomes automatic. One can often learn much more by following the thought processes of users rather than experts.

The last methodology is field testing. Observing a system deployed in the field offers a researcher less control of the variables and less influence on the task, but it makes it possible to get information and insights that cannot be gained in any other way. However, field testing is expensive and difficult to perform, and its lessons may not generalize to all communities.

Training Evaluation

The second basic approach to evaluating human factors products focuses not so much on the interaction of system and user as on the effectiveness of a particular training approach for a given system. Most training evaluations follow a five-level model called Kirkpatrick's typology, which has been in use since the 1950s (see Figure 3-3).

The first level is a simple measurement of trainees' reactions: Did they like the training? What do they plan to do with what they have learned?

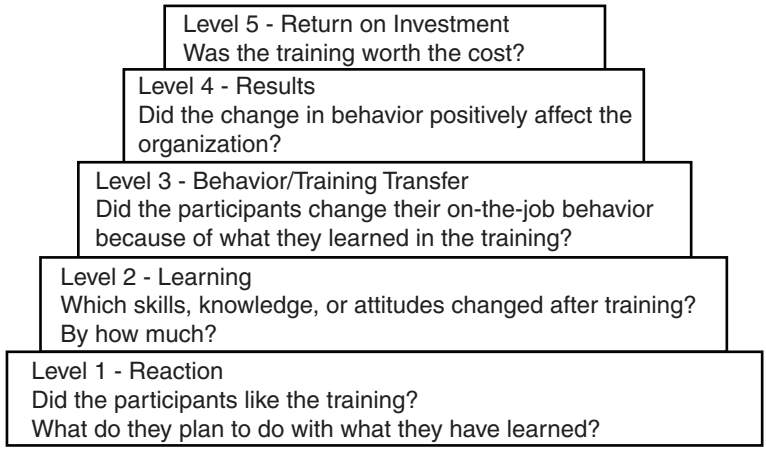

FIGURE 3-3 Kirkpatrick's model of training evaluation.
SOURCE: Kirkpatrick (1994). Reprinted with permission of the publisher. From *Evaluating training programs: The four levels*, copyright © 1994 by D.L. Kirkpatrick, Berrett-Koehler, Publishers, Inc., San Francisco, CA. All rights reserved. www.bkconnection.com.

It is the easiest and least expensive to collect, since it is basically reaction data pulled from people after a training session. "The problem is that several meta-analyses have shown the industry stops there," Salas said. A lot of people feel that if the user likes the device and the training, that's good enough, and it's okay to go ahead with putting the device in the field. The problem is that research has shown that there is very little correlation between the positive reactions of people to training and how much they have actually learned. So the fact that someone liked the instructor, liked the setting, and liked the material does not mean that the person got any real benefit from the training.

Thus it is important to move to Level 2, which measures learning. Which skills, knowledge, or attitudes changed after training? By how much? This is usually measured by a multiple-choice or some other type of test. It costs more than the first level and takes more time, but industry is moving in this direction because the evidence-based movement has convinced people of the value of getting actual data.

Level 3 focuses on behavioral changes: Did the participants change their on-the-job behavior because of what they learned in the training? This is the whole point of the training, Salas notes—people are supposed to take what they have learned and transfer it to the performance of their

jobs. Measuring such behavioral changes is significantly more expensive than measuring the acquisition of new knowledge, but it is important for organizations to know.

Level 4 looks even further and focuses on results: Did the change in behavior positively affect the organization? In a hospital, for example, the question might be whether patients are safer because of the training the hospital staff has received. Behavioral changes do not matter if they don't have positive results for the organization.

In the airline industry, for example, all flight crews have to be given teamwork training on the theory that this will lead to better performance and, ultimately, fewer accidents. And simulations do indeed show that this crew training has the desired results. But after 30 years of such training, Salas said, no correlation has been found between the training and reduced accidents. Still, the Federal Aviation Administration mandates the training.

Level 5 looks at return on investment: Was the training worth the cost? In the discussion session following his presentation, he described a Level 5 evaluation he had done for the financial firm UBS. He used a methodology called utility analysis, which, if followed systematically, makes it possible to obtain a dollar value for an intervention. However, it is based on a number of assumptions, one of which is that it is possible to assign a value to a person's performance by using interviews with supervisors. And generally speaking, Salas said, Level 5 analyses in industry are done very poorly. "It is done basically to satisfy the bean counters."

Despite the general weakness of the Level 5 evaluations, Salas said that people in the area of human factors and organizational psychology generally have a good sense of what is required for effective training evaluations. It is a very robust approach, it is systematic, it is very diagnostic, and it identifies which training works, which doesn't, and what can be done about it.

LESSONS

After the presentations of the five case studies, the presenters and other workshop participants spent some time discussing the broader lessons to be learned from these examples. Salas in particular provided some lessons from operations research.

First, Salas said, one very important fact about operations research is that, after decades of development, scientists know how to validate systems and strategies. There are still problems that can affect validation, such as the lack of leadership or bureaucratic issues, but the field has developed a set of methodologies that are theoretically driven, practical, relevant—and that work. "They are not perfect, I don't think they will ever

be perfect, but I think we know how to validate systems when humans [are] in the loop."

Twenty-five years of experience have taught him that for field evaluations of any kind to work, five things are needed, Salas said. The first is a mandate or a champion or leadership that recognizes the importance of what is being done. Evaluations take a long time, so a stable base to work from is important.

The second requirement is resources. Without money and workers, nothing can get done.

The third requirement is access to subject-matter experts, the people most knowledgeable about what is being evaluated. The problem is that these people tend to be very busy, and few of them want to spend their time talking about what they're doing rather than getting it done.

The fourth is metrics. It is vital to be able to measure the key factors in a system.

Finally, science. It is important to know how to perform evaluations in the most scientifically valid way possible.

Another lesson, Salas said, is that nobody likes evaluations. The reason is the common perception that an evaluation is looking for bad news. People will act as though they are in favor of an evaluation but then not cooperate fully when it starts. "We need a culture that can accept poor results and do something about it" rather than resist the people who are coming up with the results.

Lum focused on the importance of a research infrastructure that can be used to support the use and generation of evaluations. If such an infrastructure already exists for a field, it is something to capitalize on and take advantage of. "In counterterrorism, for example, there isn't that much evaluation research, but Christopher Koper (from the Police Executive Research Forum) and I suggest putting all the criminal justice studies into the Crime Prevention Matrix that we developed for policing and trying to glean from it some generalizations that might be applicable to counterterrorism," she said. "It is not the best, but it is all we have. So we are trying to build on something, some knowledge, in order to come to some conclusion about what might work in different areas."

If there is not already an infrastructure in a field, creating such an infrastructure should be a focus. Without one, it will be difficult to develop a system of field evaluations.

In response to a question from Robert Fein, Lum said that the NIJ had moved away from its earlier focus on evaluations.[3] The institute, particularly its science and technology division, is more focused on process evaluations and technology efficiencies. For example, in looking at

[3]For a recent review of NIJ research, see National Research Council (2010).

license plate readers, the main focus has been on the question of how well they work in reading license plates. There is still little understanding of whether the technology will help reduce auto theft, which was the major purpose of the technology.

In responding, Fein commented that the leadership of any institute is critical in terms of determining what it actually does. It appears to some that NIJ shifted quite dramatically from the late 1970s to the present in terms of a focus on evaluation. Yes, Lum said, there was a time when the institute was pushing for randomized controlled experiments, and then there was a time when that was not the case. Recently has there been a return to more evidence-based practices, under the current administration, and evaluation is getting much more attention.

4

Experiences in Other Countries

Two presentations on the second morning of the workshop offered an international perspective on field evaluation of behavioral science methods for use in intelligence and counterintelligence.

A UNITED KINGDOM PERSPECTIVE

The session's first presentation featured George Brander of the UK Ministry of Defence, who described his work with a human factors team that uses behavioral science methods to provide support to information operations. The work, he said, requires the full spectrum of human and behavioral sciences, from psychology and sociology to anthropology and even market research, and its complexity, messiness, and incomplete data make traditional evaluation and validation techniques problematic. However, he added, it is still possible to advance the state of the field through maintaining best practices.

Some 15 years ago, Brander said, he was doing traditional human factors engineering, trying to understand and develop techniques to improve the performance of UK military personnel. But about 12 years ago he found the emphasis shifting, and now his focus is on their adversaries and potential adversaries within theatres of operation. The idea was that if researchers understood how to improve their own side's performance and effectiveness, it should be possible to turn some of those techniques around and to try to reduce the performance and effectiveness of adversaries or shape the perceptions of other influential figures.

As context, Brander quoted British General Sir Michael Jackson, who said, "Fighting battles is not about territory; it is about people, attitudes, and perceptions. The battleground is there."

In their work, Brander said, there are a variety of possible foci. The first and most obvious is key individuals, such as political leaders, military leaders, business leaders, or opinion leaders. Beyond the individual level, the focus may be on teams or groups of people or larger social groups. In analyzing people at these various levels of aggregation, one can focus on such things as attitudes and opinions, cultural contexts, or the information environment in which the people function. Each of these foci requires expertise of a different sort—psychology for the study of individuals, social psychology for the study of groups, anthropology for the study of cultural contexts, market research for the study of attitudes and opinions, and so on.

Brander, who is a psychologist, initially worked with other psychologists. "Then we realized that wasn't enough," he said, "so we started to recruit anthropologists to work with us to better understand the cultural context. And then because of the importance of the information environment, we incorporated skills from media and marketing and journalism." As a way of encapsulating what his group does and where the various difficulties arise, Brander displayed a pyramid (see Figure 4-1) adapted from that originally used by Sherman Kent, often described as the father of intelligence analysis. At the bottom of the pyramid is data. The difficulties facing analysts at this level generally arise from data that are incomplete, missing, or deceptive. The middle of the pyramid represents analysis, which can be weakened by bias or flawed analytical processes. At the top of the pyramid is the answer, or the assessment together with associated "likelihood" and "confidence" levels. There are a variety of ways to evaluate and to strengthen each of the parts of the pyramid, Brander said. In the case of data, for instance, there has actually been very little work done on the validity of data, he said, and thus there is generally the possibility that a collection of data is biased in some way. On the other hand, there has been some interesting work on how to improve data, and Brander offered an example from the field of social network analysis.

The network analysis involved groups of people believed to be adversaries, he said, although he would not be more specific. The groups were being viewed as a military organization within which there were various commanders who exercised military command and control. Brander's hypothesis was that the data might also include some people who were not military commanders as such but rather who helped broker or facilitate between different organizations.

To illustrate their analysis of the data, Brander exhibited a figure that summarized a year's worth of data they had investigated (see Figure 4-2).

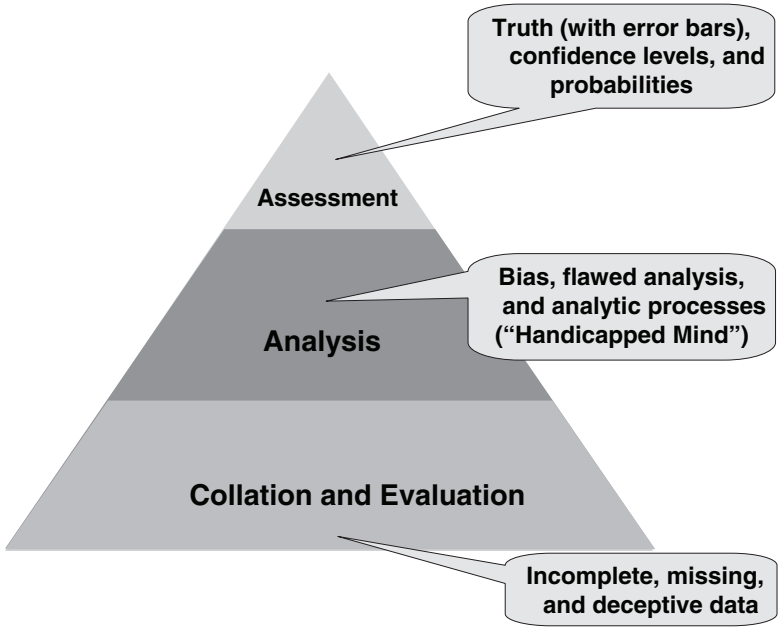

FIGURE 4-1 Analysis pyramid (adapted from Sherman Kent's pyramid).
SOURCE: Brander (2009). Reprinted with permission.

Each of the squares or diamonds on the figure represents an individual in the network, and the position of each symbol signifies both the amount of data collected on the person (on the horizontal axis) and the importance of the person as assessed by a social network analysis metric (on the vertical axis).

There were three entities in particular for which there was a great deal of data collected; the reporting for that year seemed biased toward those three entities.

In addition to noting the amount of data collected for each of the entities, Brander's group had also used social network analysis to provide a centrality measure that reflected the importance of a given person in facilitating across different networks. That was one of the particular characteristics that Brander's group was interested in. And when they performed that analysis, they discovered four individuals for whom there was relatively little data collection but who appeared to be important facilitators according to the analysis. Thus, Brander said, the analysis enabled them to overcome some of the biases in the data collection and identify people who were of interest but who would not have stood out merely in terms of the amount of data collected. Furthermore, he con-

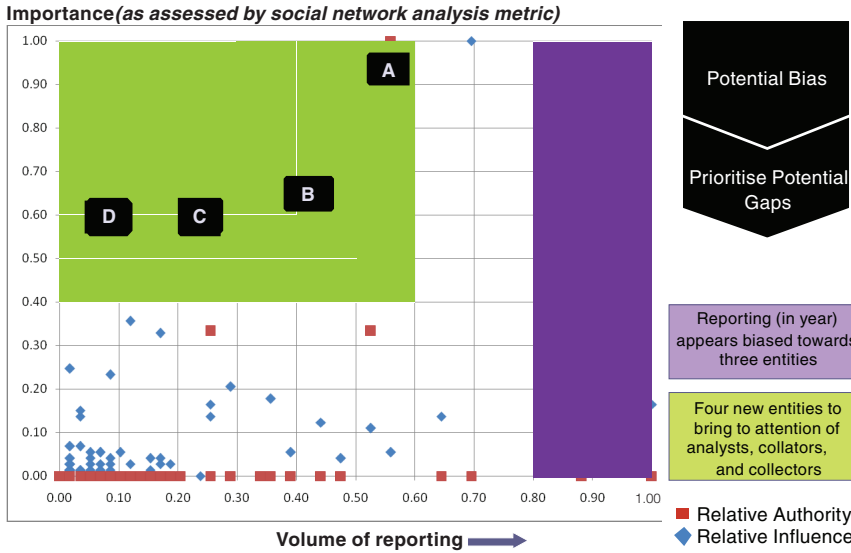

FIGURE 4-2 Identifying bias and gaps in data.
SOURCE: Brander (2009). Reprinted with permission.

cluded, the analysis did indeed identify an individual who would later become a significant player.

The case study also exemplifies the interplay between data and theory, Brander noted. Generally speaking, one collects data for a particular reason, and that reason should be taken into account during the data collection. In this case, the group was interested in individuals who were playing a role as brokers between different organizations, and social network analysis allowed them to identify such individuals who would not otherwise have been noticed. In turn, the group could point these individuals out to the data collectors and ask them to collect more data on them in order to overcome the potential bias in the initial data.

In general, Brander said, there are a variety of problems with the data available. First, the available data vary in their accuracy and completeness. Some data are incomplete, others are biased, still others have errors or may be based on some sort of deception. The data also vary according to the individuals about whom they are being collected, since people themselves vary in many different ways—age, educational background, cultural norms, motivation, access to resources, life experiences, health, and so forth. "We are trying to better understand people, and the data we get vary every time we look at a different individual," Brander said. "Having a single method is difficult."

In order to improve the data, Brander and his group have worked with the data collectors in a variety of ways, such as urging them not to throw any data away. Data collectors who may have observed such things, for instance, sometimes do not report mood or emotion for the individual concerned, so Brander's team has let them know that such information can indeed be helpful and has asked them to report it when possible.

They also have explored alternative methods of data collection, such as body movement analysis. A video of someone, for instance, can provide "a source of triangulation" that might help confirm a hypothesis arrived at through different methods.

They also work with third parties who might have observed a key individual and could report how that person responded to different things in meetings, how that person treated other people, and similar information, all of which can be used to inform the group's assessments.

Because they are operating in the context of highly variable data, individual variability, and differing cultural norms, the group has borrowed from a variety of theories and approaches to inform their analyses of individuals. These include theories of motivational style, leadership style, personality traits, and life stages. "We put in other theories because all these factors might be important to the individual we are interested in."

They also employ a variety of tools and methods: questionnaires, frameworks used to help other people make observations, content analysis of speeches, and many others. Assessing these tools and methods also demands a variety of approaches. In some cases, such as self-report questionnaires, they have a great deal of control over the method and it is relatively easy to validate. In the case of remote assessments, which makes up the bulk of what they do, it is much more difficult. They try to compare results from a variety of different methods—observed behavior, third-party assessments, case history analysis, and so forth—and see if they match up. They also use peer review to get a fresh assessment from the outside.

In the case of analysis and assessment of data, process issues are key, Brander said. There is common training for the analysts, who are generally psychologists or anthropologists who have come from the research community. The analysts are given continuous training. "We use challenge functions and discussion peer review, formal review, logbooks, and table review," Brander said. "We try and track how our methods evolve over time and problems people have had with them so we can course-correct as we go along and evolve our methods."

They also share approaches across the UK government behavioral science community as well as with their colleagues in allied countries, including the United States and various NATO countries. They have

commissioned various research studies, such as looking at alternative approaches to validation, the goal being to improve analysis and assessment of data by improving the process, even despite the limited opportunities for evaluation.

At the top of the pyramid is assssment, the outcome of the data collection and analysis. The biggest problem is determining what the customers actually want and need.

In trying to deal effectively with the customers, they keep five factors in mind: awareness, plausibility, credibility, trustworthiness, and insight. Awareness is the case of whether customers know they can ask for a particular thing. Do they know, for example, that they could come and talk to an anthropologist about how tribal dynamics work in a particular culture? Plausibility refers to whether the answers make sense to the customer. Credibility refers to whether other people agree with the answers. Trustworthiness depends on the background and credentials of the analysts. And insight refers to the implications for a decision maker—passing the "so what?" test.

Over time, Brander's group has found a variety of approaches that increase the chances of dealing successfully with customers. "We try and avoid psychological jargon because that creates all kinds of problems. We manage expectations. We say we are not predicting—we are forecasting. . . . We seek feedback on accuracy and utility; although we don't often get it. People say that was great but not much more." And they themselves try to assess the accuracy of their predictions, but it is usually not an easy task. "In terms of how a political situation may evolve or how social change may occur in Afghanistan, for example, the measures are not very good."

In summary, Brander said, much of what they do is qualitative rather than quantitative. More generally, human factors as applied to analysis and assessment is inevitably largely qualitative, so the question is whether the same quantitative approaches to validation apply. "We use multidisciplinary, multimethodologies that seek to provide insight. We try and create a sufficient degree of rigor and we try to involve best practice. We can't actually validate our tools and techniques in the traditional sense." Instead, the validation, such as it is, is done through study approaches and organizational learning aimed at helping them evolve over time toward best practice.

He quoted the English industrialist William Hesketh Lever, who once said, "Half the money I spend on advertising is wasted, and the trouble is I don't know which half." That well summarizes, Brander said, the problems he and his colleagues have with evaluating outcomes. "Some of it works. Which part we are not entirely sure."

CANADIAN DEFENSE VALIDATION EFFORTS

In the session's second presentation, David Mandel discussed his experience with Defence Research and Development Canada (DRDC), where he is a senior defense scientist and group leader of the Thinking, Risk, and Intelligence Group (TRIG) in the Adversarial Intent Section, the DRDC's human effectiveness center. An adjunct professor of psychology at the University of Toronto as well, Mandel studies various aspects of human judgment and decision making, particularly expert judgment in the area of intelligence analysis.

Mandel was put in charge of TRIG in January 2008 to carry out research related to various topics in the intelligence field. At the time of the workshop, he had hired three other behavioral scientists to work with him and was hoping to hire another soon. TRIG is currently working on three projects: one on radicalization and the economic crisis; one on developing models of state instability and conflict, which builds on the work of the political instability task force; and one on understanding and augmenting human capabilities for intelligence analysis.

Early on, he said, he found out that many people in the intelligence community with whom he was working did not understand what he did. When he first began speaking with the members from Chief of Defence Intelligence, which is the military intelligence organization that now funds two current TRIG projects, he found that he needed to establish "role clarity" concerning the functions that a behavioral science team might carry out. Some intelligence personnel thought his team might be providing behavioral analysis that would augment the agency's analytical capabilities. "I had to be clear that we were not analysts and we were not providing analytic products," Mandel said, "and I explained what we want to do really is to analyze the analytic process and to make recommendations for how to improve that."

It was a great opportunity for him, Mandel said, because as a behavior decision researcher it was very valuable to get a chance to see what analysts actually do in performing their jobs. He did find, however, that he had to be willing to switch gears from the theory-driven approach that is typical in academia to a mindset in which he paid attention to the issues that were important to those in the intelligence community and worked on those problems.

One of the challenges he has faced in working on those problems is the choice of subjects. The intelligence community tends to be skeptical of research that is done with university students, he said, and, indeed, any research that is not conducted on intelligence personnel may simply be disregarded. Yet it is difficult to free up analysts' time enough to be able to conduct behavioral studies with them as subjects.

He has come up with two solutions to this catch-22. First, he does

research on real judgments that have already been made. Instead of taking time away from analysts, he works with archival data on intelligence estimates. An advantage of this approach is that it has 100 percent external validity—that is, the results of this research clearly apply to intelligence analysts. However, the internal validity is usually lower than in experiments that are completed under the experimenter's control. Thus the research may not allow firm conclusions to be drawn about cause-effect relationships.

Mandel's second approach has been to work with trainees at the Canadian School for Military Intelligence. It does use some of the trainees' time, but the trainers are generally happy to work with the group because they are interested in the issues the group is examining. Indeed, some of the issues and training protocols that the group has studied have led to changes in the school's curriculum. In one study, for instance, Mandel and other TRIG scientists taught the trainees an analytical method based on Bayesian reasoning. The group observed a significant increase in the accuracy and logical coherence of the trainees' judgments after just a brief training.

To conclude, Mandel offered two key lessons for building a partnership between behavioral scientists and the intelligence community. First, never underestimate the importance of being poised to capitalize on opportunities. One must be ready to take action when conditions are amenable to transformative change. He illustrated this by discussing the creation of TRIG.

That creation did not come about through a top-down initiative, with senior management deciding that this particular research capability needed to be developed. Instead, it came about because of an opportunity. The chief executive of DRDC had broadened its mandate from research and development in support of defense activities to research and development in support of defense and security activities. At the same time, the center was going through an organizational realignment, which allowed Mandel the chance to describe some of his ideas to people higher up in the organization. Upper-level buy-in is important, he said, although it is also important for people at the lower levels to scan for opportunities and take advantage of them when they present themselves.

The second key lesson, he said, is never to underestimate the importance of face-to-face interactions. In the first year that he was setting up his group, he took a break from bench research and spent a lot of time in Ottawa, the Canadian capital, meeting with as many people from the intelligence community as he could—directors, analysts, trainers, administrators, and so—in an effort to understand their interests and concerns. And once he had hired the members of his team, he encouraged them to meet people from the intelligence community in order to develop famil-

iarity with them and what they did. This was particularly valuable, Mandel said, because it led the team members to become personally interested in what the analysts did and were trying to do. "We weren't just reading about a set of problems in a published document," Mandel said. "We were meeting with people who were telling us, 'These are the issues we have, and, What can you do about them?'" That allowed the team members to gain a better understanding of what the real applied issues were from the perspective of the analysts and made it easier for them to assess what it was that they could offer as scientists. "That face-to-face interaction, I think, is critical."

DISCUSSION

A significant part of the discussion following the presentations centered on the qualitative focus that Brander had described as being dominant in the United Kingdom, contrasting it with the more quantitative and device-oriented approach in the United States. Robert Fein, the moderator, said that he had had the chance over the past couple of years to get to know Brander and some of his colleagues and had been impressed with how they pushed and developed qualitative methodologies. He asked Brander why the British had gone so far in the qualitative direction and how they avoided being preoccupied with the hard technology approaches to the social and behavioral sciences that are more common in the United States.

It was partly money, Brander said—qualitative approaches tend to be less expensive than quantitative ones. But it was also because the analysts wanted to get their hands dirty and feel the data for themselves. Eventually, he predicted, they will move to more technical solutions. "For example, we are looking at ways of enhancing content analysis by some machine activity, but we wanted to know what it was like to do it ourselves before we invested heavily in the machinery."

On a related note, Brander said that the behavioral scientists and the analysts in his department tend to view technology in very different ways. He mentioned in particular a data-mining laboratory with a variety of tools to extract meaning from large amounts of data. The analysts, he said, are generally not interested unless the technology can give them more data or answers. Their attitude is, "I am just going to read everything because somebody is going to ask me a question." By contrast, the people with science backgrounds are more interested in how the technology might help them better understand their problems or how they might apply it to do better content analysis. In short, different people in the intelligence community see the advantages of technology in different ways.

Philip Rubin commented that it is important to think carefully about

technologies before they are developed. People often get caught up in the details of how the technologies work, but the more important questions are what is being measured, how well it is measured, what are the limitations with the data, and what the data can be used for.

Neil Thomason asked Brander what sorts of indicators exist for various social and behavioral features. Brander responded by discussing briefly what sorts of indicators one would use to measure social change. It is not obvious, and it depends in large part on the underlying theory used to understand and interpret social change. Should one use Max Weber? How about Foucault and the other postmodern French philosophers? In practice one place to start would be to look at attitudes, which can be measured with questionnaires. Behavioral changes would probably come more slowly than changes in attitudes. Such changes in a place like Afghanistan might take several generations to appear, he said. What are we going to measure now?

5

Ethical, Regulatory, and Cultural Considerations

The evaluation of devices and techniques for use in intelligence and counterintelligence inevitably involves the use of human subjects, and the use of human subjects requires that researchers follow various ethical and regulatory guidelines that depend on the details of the research. A session on the morning of the second day of the workshop was devoted to exploring the particular ethical, regulatory, and cultural issues that come into play when carrying out field evaluations.

ETHICAL CHALLENGES OF TRANSLATING RESEARCH INTO EFFECTIVE TECHNOLOGIES

Adil Shamoo, chair of the medical ethics subcommittee of the Defense Health Board[1] and editor-in-chief of *Accountability in Research*, opened the session with a discussion of ethical challenges of translating research in psychophysiology and neuroscience into technologies than can be used in intelligence and counterintelligence. Technologies based on psychophysiology include such things as the polygraph, voice stress analysis, the electrogastrogram, thermal imaging, and truth serums and narcoanalysis. The technologies based on neuroscience include functional magnetic resonance imaging, electroencephalography, positron emission tomography, and transcranial magnetic stimulation.

[1]Shamoo clarified that the comments made reflect his views alone and not those of the board.

The ethical challenges to developing and evaluating such technologies can be divided roughly into two categories, Shamoo said: those that arise during research and those related to the use of the technologies. Each area has its own particular issues and considerations that must be taken into account.

Ethical Challenges Associated with Human Research Subjects

The field of ethics relating to human research subjects has developed over the past 60 years, and much of that development was prompted by concerns over ethical lapses. For example, the Nuremberg Code was developed in 1947 to set out principles for human medical experimentation in response to what had been uncovered during the Nuremberg trials about experiments performed by Nazi doctors on Jews in concentration camps and other prisoners. The Helsinki Declaration of 1964 was produced by the World Medical Association to be a universally accepted set of ethics principles governing the behavior of doctors and other researchers doing studies with human subjects, and it included many of the same principles set out in the Nuremberg Code. In the United States the Tuskegee Syphilis Study, a controversial 40-year study of nearly 400 poor black farmers with syphilis, led to the establishment in the early 1980s of regulations to protect human subjects and later to the creation of the Office of Protections from Research Risks and to the requirement that federally funded human subjects research be overseen by institutional review boards.

The spirit of all these ethical guidelines was captured in the words of an 85-year-old survivor of the Nazi concentration camps, Shamoo said. Eva Mozes Kor and her identical twin sister were subjects of the experiments that Josef Mengele performed on Jewish concentration camp prisoners during World War II. "They both survived, but they went through several months of hell, and they have come to this country, and she lives now in Terra Haute, Indiana." Once, Shamoo said, Kor had been invited to talk to a meeting of about 3,000 doctors and medical researchers, and her words remained with him. "She said, 'You, the scientists of the world, must remember that research is done for the sake of mankind and not for the sake of science.'"

Although there are a variety of regulations covering various areas of research, they are all attempting to formalize the behavior that Kor was advocating: that researchers always remember that their work is done to benefit mankind, not simply to advance science. And as such, Shamoo said, the responsible conduct of research can be encapsulated in a few basic principles.

The first is honesty. Easy to say, easy to understand, but not always easy to adhere to.

The second is objectivity. Shamoo says that he gives his students a list of 20 steps in conducting research, from forming a hypothesis and doing a literature review to the collection and analysis of the data and publication, and he asks them in which steps they could bias the outcome. "They all pick one or two," Shamoo said, "and usually there is one very smart student who says in every one of these steps, which is true." Objectivity requires doing nothing in any of these steps to bias the outcome.

As an example of how such bias can creep into a study, Shamoo mentioned the case of Viagra. It is marketed on television to men in their sixties and seventies because those are the people most likely to use the drug. But the original clinical trial was conducted on a population whose average age was 56, raising questions about the applicability of the trial to the market audience.

The third principle is respect for research subjects, and such respect demands that a research project meet a number of criteria. It should be scientifically valid, for if it is not then the research subjects are risking potential harm for no potential gain. It should have social value and be beneficial to individuals or to the larger society in some way. The researcher should obtain informed consent from all of the subjects. That is, each subject must understand the purpose of the research, how it will be conducted, the possible risks to the subject, and the potential benefits to the subject and, based on that understanding, must voluntarily agree to take part. The subjects should be selected equitably, and no potential subjects should be taken advantage of simply because they are easily available. The more vulnerable subjects are, the more they should be protected. And there should be independent review of the research, such as by institutional review boards.

A few years ago, the National Research Council released a report, *The Polygraph and Lie Detection* (National Research Council, 2003). One of its recommendations was that any research in this area should follow "accepted standards for scientific research" and should "use rules and procedures designed to eliminate biases that might influence the findings." These are two key principles, Shamoo said—following accepted standards and eliminating biases. "These are the heart and soul of responsible conduct of research."

A variety of regulations govern ethical issues dealing with human research, Shamoo noted, and which regulations apply to any particular research study depends on who is funding the study. If a study is funded by the federal government and it involves human subjects, the researchers must follow the Common Rule (45 CFR 46). This includes research funded by the U.S. Department of Defense, the Department of Energy, the National Institutes of Health, and many other departments and agencies. The agency responsible for oversight of research involving human sub-

jects is the Office for Human Research Protections in the U.S. Department of Health and Human Services, although individual agencies have their own oversight offices.

Much of the research performed by private industry with human subjects in support of data for marketing a drug or device is regulated by the Food and Drug Administration (FDA), which also regulates the marketing of various drugs and devices for human consumption or use. Privately funded research, however, is not regulated in the United States, Shamoo noted.

To finish up the part of his presentation devoted to human subjects research, Shamoo provided some details on the FDA's regulation of drugs and devices. First, he noted, not only does the FDA follow federal regulations on human subject research, such as 20 CFR 50, which closely resembles the Common Rule, but it also has a phased approach to the development, testing, and marketing of drugs and other products, with each phase having a different set of requirements for approval. Phase I studies are initial human studies done in healthy volunteers to determine the minimal toxic dose. In Phase II studies, the drugs are used for the first time in people with the illness the drug is designed to treat; they usually involve about 20 to 80 people, continue the examination of the drug's safety, and begin to get initial data on its efficacy. Phase III studies are randomized, controlled, multicenter trials on a few hundred to a few thousand people. Once a drug is approved by the FDA and marketed, it enters Phase IV, which is postmarketing surveillance looking for unexpected side effects that might not have shown up in the earlier clinical trials.

Ethical Challenges Associated with Using Technologies

Besides ethical issues related to human subject research, a variety of ethical issues are raised when technologies move beyond the research stage and are put to general use. In particular, a number of technologies now under development could be very useful to the intelligence community, but they also raise serious ethical concerns. Jonathan Moreno, a professor of medical ethics and of the history and sociology of science at the University of Pennsylvania, described a number of these developing technologies and the sorts of issues that will need to be addressed if these technologies are to be put to work.

Two reports have appeared recently from the National Academies that discuss potential applications of neuroscience research, Moreno noted. The first one, *Emerging Cognitive Neuroscience and Related Technologies* (National Research Council, 2008),[2] has a bland title but contains

[2]Moreno was a member of the committee that prepared that report.

some very interesting ideas about how neuroscience techniques might be applied to the fields of intelligence and counterintelligence. The technique that has attracted the most interest to date is functional magnetic resonance imaging, or fMRI. It makes it possible to watch which parts of the brain are most active during different activities, which has potential for allowing researchers to determine, to at least some degree, what types of things a person is thinking about. It may be, for example, that more of the brain is activated when there is intentional deception than when one believes one is telling the truth, so some believe that fMRI or a similar imaging technology might someday serve as an accurate lie detector.

A number of other neuroimaging techniques could be used in similar ways, including positron emission tomography and near infrared spectroscopy. "I think within 10 years we will have much more granular pictures of what is going on in the brain while people are doing things or looking at things," Moreno said. "Is that mind reading? Is that brain reading? I don't know. I have my doubts."

One problem with fMRI, he noted, is that the machines are not particularly practical for use in the field because they are very heavy and also quite noisy. Some researchers have been working on the development of portable fMRI units, he said, although he did not know how close to success they are—if at all. "My guess is that it hasn't advanced very far." Another problem is that the very notion of a lie is conceptually far more complex than people ordinarily realize, which creates a fundamental obstacle to "objective" deception detection.

A second neuroscience-related area with potential applications to the fields of intelligence and counterintelligence is psychopharmacology, or the study of drugs that affect thinking, mood, and behavior. One example is the use of oxytocin, a neurotransmitter that is associated with a number of behaviors, including trust and love. It can be administered in spray form through the nose, Moreno said, and about a half-dozen studies have found some evidence that oxytocin can cause people to act in a more trusting way under experimental conditions. However, some neuroscientists do not believe that oxytocin can get past the blood-brain barrier into the brain, Moreno noted, so there is some controversy as to whether the studies are valid.

These sorts of technologies raise a variety of ethical questions that society has not yet begun to address, Moreno said. Suppose, for example, that the oxytocin research shows that it is indeed possible to get people to answer an interrogator's questions because a quick squirt of it in the nose leads them to feel as though they can trust the person asking them questions. "Would that be more acceptable than pressuring him or her through physical means," Moreno asked, "or is this going to the heart of what it is to be a human being? Does this violate cognitive privacy? I don't

know how to answer that question, but I think it is one that we will face. If not with oxytocin, then with something like it."

In his presentation, Shamoo voiced similar concerns about the use of fMRI and other neuroimaging techniques. He quoted the bioethicist George Annas as saying that these new devices are particularly threatening to individual privacy because of the potential that they could be used to peer into a person's brain with or without that person's permission. How can the privacy concerns be addressed? Do the potential benefits to society outweigh the risks to the individual? These are the sorts of questions, Shamoo said, that people must ask themselves before moving ahead with these devices.

And these are not just theoretical issues, Moreno said. Just a few months before the workshop, the National Research Council published *Opportunities in Neuroscience for Future Army Applications* (National Research Council, 2009), which discusses a number of potential technologies of these types. "These are serious scientists who think there are going to be advances that are plausible for the army to invest in during the next five or ten years."

Moreno mentioned transcranial magnetic stimulation in particular. This technique, which uses magnetic fields to induce changes in brain activity, influences such brain functions as visual perception, memory, speech, and mood, and it may have the potential to alter a person's social behavior or attitudes. One of the report's recommendations was that the army should examine transcranial magnetic stimulation for enhanced learning in soldiers.

If the army chooses to pursue such applications, Moreno pointed out, it will require extensive research and, eventually, field testing, both of which will raise ethical issues that have yet been worked through. Perhaps even more challenging will be the ethical issues associated with the widespread use of such technologies. "We have already had some preliminary experience with this with the anthrax vaccine controversy," Moreno said, referring to the controversial policy that ordered more than 200,000 soldiers during the first Gulf War to get an anthrax vaccination in case of a bioweapons attack. With more and more technologies being developed to improve the performance of soldiers, the question arises of how modified soldiers will have to be in the future. How much will society require them to accept? How much will the individual soldier accept? In developing these technologies and putting them to use, Moreno said, the researchers and others involved should be careful that it is all done with respect for the people involved and with respect for the proper ethics at each step.

FIELD TESTING VERSUS RESEARCH

When discussing ethical and regulatory issues in field evaluations, it is important to keep in mind the differences between field testing and research. As Moreno explained, the two are not identical, and they require somewhat different approaches and considerations.

"Obviously not all research is field testing," Moreno noted, "and that is illustrated by the fact that there are people in labs who do research but who are not necessarily going out in the field." This is not particularly surprising to most people. What is surprising, however, is that not all field testing is considered to be research. As an example, Moreno pointed to the more than 200,000 men who were deployed at above-ground atomic bomb tests from 1948 to 1963. Many of these men were given radiation badges that indicated levels of exposure to radiation. Some of the pilots who flew through the mushroom clouds were dusted for radioactive particles. Their urine and other bodily fluids were checked for radioactivity. Still, Moreno said, they were not considered to be human research subjects at the time, and even within the current understanding of research rules they might not be considered to be research subjects.

The reason that they were not research subjects even though scientists were able to gain a great deal of information from these activities is that they were there for other purposes. Specifically, they were there for training and for desensitization to the atomic battlefield.

The key point here, Moreno said, is that field testing that is not considered research is not subject to the various ethical and regulatory requirements that govern research. For instance, the usual research rules about informed consent and prior peer review do not apply to this kind of field testing. This doesn't mean that no ethical or regulatory standards apply, but it does mean that many of the usual requirements governing human subjects research may not apply, such as the Common Rule, FDA regulations, and certain Department of Defense regulations.

There is inevitably a certain amount of gray area between research and field testing, and this opens up the possibility of gaming the system. Moreno described a study done in the early 1990s at a hospital in New York looking at two different ways of doing sutures for face-lifts. Each of the approximately 20 face-lift patients had one type of suture done on one side of the face and the second type of suture done on the other side. The surgeons did not consider the study to be human subjects research and so did not fulfill the usual regulatory requirements, such as getting approval from an institutional review board. And it was close enough to the gray area, Moreno said, that they probably could have characterized this practice as innovative surgery, except for one thing: they published their results as a research study.

Of course, Moreno commented, the surgeons should have considered their work to be a research study from the beginning because they were going about it in a very controlled and systematic way, but the problem became very clear and public for them only when they published the results in the research section of a surgery journal.

While some field testing is not research, a great deal of it is, and the overlap between field testing and research is referred to as research field testing. It can be defined as systematic investigations that are carried out under actual field conditions.

It turns out that some of the soldiers and marines deployed near the above-ground nuclear tests actually were considered to be research subjects taking part in research field testing between 1953 and 1962. In particular, they were taking part in psychological studies known as "panic studies." The Department of Defense was concerned about how soldiers would react if they were close to an atomic explosion and what the psychological effects might be, so a group of psychologists and psychiatrists were hired to perform tests on a group of subjects before and after an atomic blast. In one test, for example, soldiers were told to disassemble and reassemble their rifles within minutes of the explosion, while the researchers observed them for signs of panic. The soldiers who took part in these studies were treated as test subjects and gave their consent to participation in the studies. Thus they were treated differently from the tens of thousands of other soldiers and marines who were near the blast sites when the bombs went off but were considered as being deployed for training exercises.

Today, by contrast, there are many field trials undertaken in hospital emergency departments around the country, such as the testing of a new method to treat heart attacks, and they are considered to be clinical trials and therefore require informed consent from the patients and prior approval by an institutional review board. An FDA rule covers these emergency medicine trials and specifies the procedures that must be followed—a situation that creates bureaucratic hurdles that frustrate many who do research in emergency medicine, Moreno said.

A key issue here—and one that is often not mentioned in the ethics literature—is who decides whether an activity is a field test or a research study. "If sending you to function within a mile or two of ground zero is not considered to be a human experiment, then informed consent does not apply," Moreno noted. "This is a key point, because it is possible to game the system." Moreno said he sees examples of this in medical fields, such as surgery. Some physicians may carry out experiments but do not characterize them as clinical trials. They keep track of the results as a series of cases and are careful not to publish the series in a journal,

which means the work may not be covered by the requirements governing clinical research.

Despite the opportunities for taking advantage of the gray area, researchers doing field testing should adhere to normal ethical and regulatory procedures, Moreno said. "Field testing that includes development, testing, and evaluation designed to develop or contribute to generalizable knowledge is subject to prevailing ethical and legal conventions governing research."

DISCUSSION

The discussion session following the two presentations expanded on the speakers' comments and introduced some new topics as well.

As Robert Fein noted, the ethical issues involved with field evaluation sometimes come face to face with various political and economic issues. For example, the Department of Defense is interested in getting devices and other technologies to detect deception into the field as quickly as possible, and private companies have economic incentives to do the same thing. How, he asked, do these pressures interact with privacy and individual rights concerns in field evaluation?

"The answer," Moreno said "is that in a pinch there is a tendency for the bar to be lowered because of political pressure and legitimate public concern about taking care of our men and women." As an example, he mentioned the drug pyridostigmine bromide (PB), which was given to troops in the first Gulf War in case of exposure to the nerve agent soman; the drug was a pretreatment that would improve the effectiveness of the treatment for soman exposure, but at the time it had not received FDA approval for medical use. Later, many alleged exposure to PB and other drugs to be associated with the development of various health problems in Gulf War veterans. When members of congress asked the FDA why it had given the Department of Defense a waiver for the informed consent that would have normally been required to use the drug, Moreno continued, "the FDA said, we're not the war fighters. If the Defense Department comes to us and says we need to do this to protect our people, are we going to say no? We're in the business of approving drugs and devices for medicine, not for fighting a war."

The situation was similar with the anthrax vaccine given to Gulf War soldiers. Some of the soldiers later blamed the vaccine for various health problems that were part of the Gulf War syndrome, and they complained that they felt like human guinea pigs, given something without consent. From the defense department's point of view, the move was necessary to save lives—potentially thousands of lives—and to maintain force readi-

ness. At the same time, Moreno said, if there is not transparency, if there is not public confidence in a decision, then people can end up feeling as if they were exploited.

The bottom line, he said, is that the bar is naturally set lower in those situations in which the use of a drug or technology still being tested could save the lives of many soldiers or other people defending their country. In such cases, the tendency is to loosen the ethical restrictions somewhat.

Shamoo noted, however, that in response to the experience with PB, the FDA is now required to get approval from the president to bypass its usual regulations, as was done in that case. In the future, only the president will have the power to loosen the guidelines, even in the most pressing cases.

On the issue of what constitutes research versus simple surveillance, planning committee member Robert Boruch of the University of Pennsylvania noted that the level of record-keeping seems to play a large role. He mentioned a recent case involving researchers from a top-tier university who conducted a randomized trial in hospitals in which doctors and other health care providers were encouraged to wash their hands and engage in a series of other check-listed activities to enhance hygiene and reduce infection. The researchers did not seek permission from an institutional review board, which led to the university being sanctioned by the Office for Human Research Protections. That office judged that the trial was actually an experiment because it was an effort to systematically understand the extent to which hospitals could get health care providers to be more conscientious about hygiene for the sake of their patients and to estimate the effect of the effort on such things as infection rates.

Moreno responded that deciding what research is can be a difficult problem. Suppose, for example, that a researcher approached a nursing home with a project to convince staff members to wash their hands between patient encounters in order to avoid bacterial infection. The researcher might even help them develop a program to increase hand washing based on some information that the researcher had gathered about the employees' baseline hand-washing practices. If the researcher then published the results, that might suggest that it had been a research study for which approval from an institutional review board was required and to which the patients would have had to give their consent. But not necessarily—it might still be considered a hygiene program that got reported as a case study. But if the researcher then went on to compare the program with educational hand-washing programs in other nursing homes, then the program moves closer to being a research study. "It's not an easy line to draw," he said, "but I think you can intuit those lines."

Christian Meissner commented that, from his experience as chair of an institutional review board, he knows that there is a significant gray

area between program evaluation and research. Indeed, he said, it is quite possible to field test things under the guise of program evaluation. But once one begins manipulating factors and having control groups, the studies clearly amount to research.

David Mandel touched on a similar issue. In his studies of analysts and analyst trainees, he often deals with research that has begun years earlier. In one particular study of the calibration of intelligence estimates, he was dealing with estimates generated years before his research group was even created. The agency that produced the estimates was not interested in research ethics issues, he said—from the agency's perspective, it was a quality control exercise rather than an example of research, and they had been going about it long before they came to think of it as research as a result of their partnership with Mandel and his team. From Mandel's perspective, however, he was engaged in research and was bound to go through the institutional review board process, even though the research had already been done. "Once it moves from an internal quality control exercise to a collaboration that has a research side to it," he said, "we have to put it through our IRB even if we can't go back and get informed consent because some of those analysts have moved on." And, as the research moves forward, Mandel's group has to get consent forms from the analysts in order to use their assessments for research purposes. Sometimes the same exercise is both research and not research, and in those cases Mandel's group—as researchers—must treat the work as research and follow the standard research procedures.

6

Looking to the Future

Over the two days of the workshop, a significant portion of the presentations plus a large percentage of the discussion focused on the future. Presenters and participants talked about the obstacles to field evaluation of techniques derived from the behavioral sciences and intended for use by the intelligence community, about general lessons from other fields about what it takes to implement field evaluations in a serious and comprehensive way, and about some of the particular implementation issues in the intelligence arena. The discussions were realistic about the obstacles but optimistic about the possibility of eventually developing a culture within the intelligence community in which field evaluation is accepted as a necessary and usual feature. The discussions also included a focus on the best path forward.

OBSTACLES TO FIELD EVALUATION

In one of the discussion periods, Neil Thomason commented that he had been struck by the difference in testing and evaluation between law enforcement and the intelligence community. Christian Meissner had identified many hundreds of research papers from the past several decades that applied to eyewitness identification, Thomason noted, while Thomason himself had been able to identify only six papers on the Analysis of Competing Hypotheses (ACH) from the same period. "It is just two totally different worlds," he said. But why should this be, he asked. Why is it that when a technique or a device is developed for use by the intel-

ligence community, there is so little attempt to evaluate it in the field to see if it really works?

It is particularly puzzling, he said, in light of a comment by Steven Kleinman, who had suggested that one of the weaknesses of the American intelligence community is that it has too much money. Because so much money is thrown at intelligence work, he said, "there is a built-in assumption that if we don't get it right, somebody else will." If the HUMINT (human intelligence) groups don't figure something out, then the SIGINT (signal intelligence) people will, and if SIGINT doesn't get it, then IMINT (imagery intelligence) will. But why, Kleinman asked, hasn't more of this money been used for field evaluation studies?

A number of the workshop presenters and participants spoke about various obstacles to field evaluation inside the intelligence community—obstacles they believe must be overcome if field evaluation of techniques and devices derived from the behavioral sciences is to become more common and accepted.

Lack of Appreciation of the Value of Field Evaluations

Perhaps the most basic obstacle is simply a lack of appreciation among many of those in the intelligence community for the value of objective field evaluations and how inaccurate informal "lessons learned" approaches to field evaluation can be. Paul Lehner of the MITRE Corporation made this point, for instance, when he noted that after the 9/11 attacks on the World Trade Center there was a great sense of urgency to develop new and better ways to gather and analyze intelligence information—but there was no corresponding urgency to evaluate the various approaches to determine what really works and what doesn't.

David Mandel commented that this is simply not a way of thinking that the intelligence community is familiar with. People in the intelligence and defense communities are accustomed to investing in devices, like a voice stress analyzer, or techniques, such as ACH, but the idea of field evaluation as a deliverable is foreign to most of them. Mandel described conversations he had with a military research board in which he explained the idea of doing research on methods in order to determine their effectiveness. "The ideas had never been presented to the board," he said. "They use ACH, but they had never heard of such a thing as research on the effectiveness of ACH." The money was there, however, and once the leaders of the organization understood the value of the sort of research that Mandel does, he was given ample funding to pursue his studies.

One of the audience members, Hal Arkes of Ohio State University, made a similar point when he said that the lack of a scientific background among many of the staff of executive agencies is a serious problem. "If we

have recommendations that we think are scientifically valid or if there are tests done that show method A is better than method B, a big communication need is still at hand," he said. "We have to convince the people who make the decisions that the recommendations that we make are scientific and therefore are based on things that are better than their intuition, or better than the anecdote that they heard last Thursday evening over a cocktail."

A Sense of Urgency to Use Applications

A number of people throughout the meeting spoke about the pressures to use new devices and techniques once they become available because lives are at stake. For example, Anthony Veney spoke passionately about the people on the front lines in Iraq and Afghanistan who need help now to prevent the violence and killings that are going on. But, as other speakers noted, this sense of urgency can lead to pressure to use available tools before they are evaluated—and even to ignoring the results of evaluations if they disagree with the users' conviction that the tools are useful.

Robert Fein described a relevant experience with polygraphs. The National Research Council had completed its study on polygraphs, which basically concluded that the machines have very limited usefulness for personnel security evaluations, and the findings were being presented in a briefing (National Research Council, 2003). It was obvious, Fein said, that a number of the audience members were becoming increasingly upset. "Finally, one gentleman raised his hand in some degree of agitation, got up and said, 'Listen, the research suggests that psychological tests don't work, the research suggests that background investigations don't work, the research suggests interviews don't work. If you take the polygraph away, we've got nothing.'" A year and a half later, Fein said, he attended a meeting of persons and organizations concerned with credibility assessment, at which one security agency after another described how they were still using polygraph testing for personnel security evaluations as often as ever. It seemed likely, Fein concluded, that the meticulously performed study by the National Research Council had had essentially no effect on how often polygraphs were used for personnel security.

The reason, suggested Susan Brandon, is that people want to have some method or device that they can use, and they are not likely to be willing to give up a tool that they perceive as useful and that is already in hand if there is nothing to replace it. This was probably the case, she said, when the U.S. Department of Defense (DoD) decided to stop using voice stress analysis–based technologies because the data showed that they were ineffective. The user community had thought they were useful, and when they were taken away, a vacuum was left. The users of these

technologies then looked around for replacement tools. The problem, Brandon said, is that the things that get sucked into this vacuum may be worse than what they were replacing. So those doing field evaluations must think carefully about what options they can offer the user community to replace a tool that is found ineffective.

Philip Rubin offered a similar thought. The people in the field often do not want to wait for further research and evaluation once a technology is available, he said, and "there are those out there that will exploit some of these gray areas and faults and will try to sell snake oil to us." The question is, How to push back? How to prevent the use of technology that has not been validated, given the sense of urgency in the intelligence field? And how does one get people in the field to understand the importance of validation in the first place? These are major concerns, he said.

Institutional Biases

Some of the most intractable obstacles to performing field evaluations of intelligence methods are institutional biases. Because these can arise even when everyone is trying to do the right thing, such biases can be particularly difficult to overcome.

Paul Lehner began his talk with a story about field evaluation that illustrated how such biases can come into play. He had been involved in a study that evaluated how much analysts should rely on a certain type of information that they use fairly routinely. He and his colleagues had developed a simple method for retrospectively evaluating the accuracy and value of the information that the analysts were using, and they compared that retrospectively analyzed value with what the analysts had been told at the time about the value and accuracy of the information.

Their results indicated that the system being used to evaluate the information the analysts were getting was very inaccurate. Indeed, according to their study, information that was thought to be of less value was seen retrospectively as being substantially more accurate than information that had been labeled as having higher accuracy.

It was a small study, so it could not be definitive, but the important fact was that the study was easy to do and could have been repeated half a dozen times for probably less than a year of staff time, and then the results most likely would have been definitive. But that never got done.

The original sponsor who had championed the study had moved on to a new position. The new sponsor saw that the results ran counter to conventional wisdom and decided not to release the study until it had been reviewed. So the study was sent out for review—to the organization that created the particular sort of information that was the subject of the

study. This made sense, Lehner notes, since the people of that organization were the experts on the subject. But the senior expert in that organization did not believe the results and so never responded to the request for release. That made sense as well, Lehner said. "If I was that person, I would probably do the same thing. I would never say go ahead and release it, because clearly the results were wrong. Also, I would never send a formal reply recommending that the study not be released, because then I would be on record for suppressing a negative study." So the smart thing to do was simply not to respond, which is what happened. As a result, the study was never published, and no one else ever got to see it.

This is a common way that things can go wrong with a field evaluation, Lehner said. He had experienced the same thing in slightly different versions three times in the previous six years.

What went wrong? A number of factors combine to produce this sort of situation, Lehner said. The first factor is the requirement in the intelligence community to get permission for anything you want to do. This makes sense, given that the release of the wrong information could result in people getting killed, but it creates a situation in which it is easy for information to be suppressed.

A second factor is practitioner overconfidence. People tend to have confidence in the tools and methods they have experience with and to believe that their own experience is more trustworthy than the results of a researcher who comes into an area and conducts experiments.

The third factor is organizational and bureaucratic. Field research generally requires a champion to obtain the funding and pave the way politically, but senior people tend to move around a great deal in bureaucracies, and the chances are that the champion will have been reassigned before the study is complete. The new manager is unlikely to push for—or even believe—the study that the previous manager had championed. And so the study dies of neglect.

All of this points to a basic conclusion, Lehner said: in the intelligence community there is a strong institutional bias against obtaining or reporting negative results. The bias does not arise for political reasons or from people protecting their turf. Everybody involved is trying to do what they think is the right thing. Still, the combination of factors creates a situation in which it is very difficult to perform and report field evaluations that call into doubt methods that are being used.

Something similar happens when new techniques are introduced. The people who introduce new methodologies and tools generally believe in their practices; otherwise they would not be introducing them. So most of these people believe that if a good field evaluation were to be performed, the particular methods they are introducing would pass. A corollary is that if these people are given the choice between putting their method into prac-

tice or waiting until a field evaluation is performed, they would generally go ahead. Why wait when you're sure it works?

But that leads to a problem. Once the new method has been put into practice, there are now people who are experienced with it and are certain that it works. No matter how good or bad it is, there will be at least some experiences in which everything works out well and the practitioner now has faith in the method. As Lehner phrased it, "It becomes part of the tried-and-true methods."

The workshop had already provided a couple of examples of this pattern, Lehner noted. As Thomason noted, the technique of ACH has achieved a cult-like status in the intelligence community without ever having had a serious field evaluation. Similarly, Veney described the Preliminary Credibility Assessment Screening System (PCASS) as a "godsend on the battlefield" even though it has never had a true field evaluation.

The main reason that such methods become part of the intelligence toolkit, Lehner said, is that they satisfy a need. New methods and tools are not put into the field because there is a great deal of evidence showing that they work. They are put into the field because something is needed to fill a void. And once they become part of the accepted set of methods, it becomes very difficult to produce negative evaluations of them, for all the reasons described above.

This in itself wouldn't be a problem if most of the new methods worked, but that is not the case, Lehner said. Even many of the ideas that are supported by validating field experiences don't work. Expert judgment and field experience are surprisingly poor at discriminating between what works and what doesn't. "You see this over and over again in lots of different fields. We see it here, too."

Lehner predicted that if the three promising methods described earlier in the workshop—ACH, PCASS, and APOLLO—were field evaluated, only one of them would pass. "I have no idea which one," he said, "because most good ideas don't work, even those supported by experience (but not objective testing). So just going with the base rates, I would guess that one of these methods works and two do not."

LESSONS FOR THE PATH FORWARD

Although there are many obstacles to reaching a point at which field evaluations are a regular and accepted part of the process of adapting techniques from the behavioral sciences for use in intelligence and counterintelligence, workshop speakers identified a number of things that can make that path easier. In particular, they accumulated a number of lessons that offer components of a potential framework for taking something from the laboratory to the field.

A Trigger

In reviewing his presentation on research into eyewitness testimony, Meissner described a number of the factors that brought the field to the point of having a wealth of research papers bearing on the issue. The first was what he termed a "key sociological event"—the DNA exonerations proving that a number of people convicted on the basis of eyewitness testimony were actually innocent. "That shocked the system," he said. "It not only spurred additional research on the part of experimental psychologists but also encouraged the system to change." In short, the DNA exonerations acted as a trigger that set a number of things in motion, including increases in funding and a heightened interest in the subject on the part of researchers.

Meissner noted that the 9/11 attacks also served as a trigger of sorts for increased interest in the issue of interrogation. He had already been doing research on interrogation in the criminal justice realm, but it was only after the attacks that funding began to be available for research on interrogation in the areas of intelligence and counterintelligence. "There were just a handful of folks doing research in this area," he said, "but now more and more researchers are coming to the table."

Funding

A second lesson is the importance of funding for field evaluations. Grover Whitehurst made the point explicitly in talking about lessons from the field of education: "We need more investment. We need fair and open ways for people to compete for the funds from those investments to create knowledge. We need to develop priorities for those investments that move the university-based research community towards questions that are important to practitioners and policy makers. Most academics want to talk to themselves, not to people in the field, and there are ways to incentivize them to move from the bench to the trench."

Meissner offered the same lesson from the area of psychology and law. "Having a mechanism that is constant, that is competitive, that is independent is really important to getting good science funded," he said. If field evaluation of techniques in intelligence and counterintelligence is to advance, it will require a steady, reliable funding stream that is structured to attract academic researchers to work with those in the field to develop a body of evidence.

A Research Base

If field evaluations are to be convincing and useful to practitioners, Meissner said, they need to be part of a larger, multimethodological

research base in which the different pieces are consistent and support each other. For example, if he and other researchers in psychology and the law had had only a few studies about eyewitness testimony, they would not have been able to convince the legal community that they needed to change. But in fact, he said, they had a very robust research literature that was both high quality and extensive. They also had a consistency of findings across different methodological approaches, using a diversity of methods and analytic approaches, which indicated a general agreement among scientists.

Basic research is an important part of it, Meissner said. The plethora of studies he mentioned include not only focused eyewitness studies but also studies that examined how memory works and how people recognize faces, models of face recognition, models of memory, models of social influence, and much else. In the intelligence area, he said, there is a great deal of basic research being done in the laboratory that on the surface doesn't seem to have any relevance for what analysts do; in fact it is highly relevant to the basic processes that influence analysts' decision making.

Finally, he said, the research on eyewitness identification also includes a strong theoretical grounding. Indeed, there are formal mathematical models of eyewitness identification that not only replicate previous work but also predict future findings.

Ongoing work on interrogation, Meissner said, is also engaging in a systematic program of research. It includes experimental laboratory studies, field research, and surveys. It includes research on experts in the art of interviewing in an attempt to determine what makes a person an effective interrogator. It is surveying the literature. And researchers are collaborating with practitioners. This is consistent with the tiered approach suggested by Charles Twardy, in which initial research might be done with psychology students and more refined testing in an intelligence academy or with working analysts.

In her talk on policing, Cynthia Lum made the point that a solid body of research is important in getting practitioners to accept and use the work. A number of police—particularly lieutenants and higher—come to her center's website and use the interactive tools to find studies that give them ideas for how to deal with particular issues.[1] The response has been very positive, she said, because many of these police officers are being pressured to say how they are going to deal with a particular crime problem and they need to be able to back up their answer with some proof that it is going to work. The collection of research studies available on Lum's site provides exactly that sort of evidence.

[1]See http://gemini.gmu.edu/cebcp/index.html.

Engagement with Practitioners

A recurrent point to emerge from the discussions at the workshop was the importance of researchers establishing and maintaining a good relationship with practitioners. Meissner stated it succinctly: "It is really important to collaborate and engage the practitioners, to bring the practitioners into the laboratory, to work with them on the very problems that you are facing, to understand the issues of implementation." This includes ensuring they understand that if methods are implemented differently than designed or adapted inappropriately, it can produce unvalidated approaches.

What are the keys to a successful engagement with practitioners? The workshop participants offered several different perspectives. The group discussed the potential value of researchers who wish to communicate well with practitioners being able to transmit information through stories. The practitioners themselves—whether intelligence analysts, police officers, or educators—tend to pass information along through stories, so if researchers are to communicate their results effectively to the practitioners, they would do well to become good storytellers.

George Brander of the UK Ministry of Defence agreed that telling stories is vitally important to practitioners. The model that has evolved in the United Kingdom, he said, is that people join the research community with skills in anthropology, psychology, sociology, or some other area of behavioral science; they start doing their research, they get closer to the practitioners and learn how best to interact with them, and eventually they figure out how to effectively provide them with advice and guidance—which often includes telling stories.

Kleinman added that storytelling is important because "there is frequently an inverse relationship between authority and expertise" in which the people who make the decisions generally will understand relatively little about the scientific details. This is why, he said, the "snake oil salesmen" are able to convince people to use techniques for which there is little or no evidence of effectiveness. They are excellent storytellers, he said. "They would have very weak data, so they don't spend much time on it, and they definitely make sure their audience is carefully selected so that people like those in this audience, who would cut them to shreds, are noticeably absent." Thus, he said, it is important for researchers to be able to step outside their normal linguistic comfort zones and communicate in the way these decision makers do—that is, with stories, clear images, and a strong focus on what is in it for them.

Heather Kelly of the American Psychological Association said that storytelling is particularly important when dealing with Congress—and it is Congress that ultimately controls what gets funded and what does not. The importance of storytelling is one important reason why it is easier to

sell applied research, such as that done by the Department of Defense, than it is to sell basic research, such as that done by the National Science Foundation. "I would like for you all to be thinking about the best stories that we can tell on Capitol Hill," she said. "It is particularly powerful when it comes from outside basic researchers versus inside researchers."

Mandel offered a different perspective, suggesting that a more important skill than storytelling for scientists is being able to listen and being open to looking at scientific issues from the point of view of the practitioners. Research scientists are generally more interested in testing theories than in examining practical problems that are of importance to the practitioner community, and the scientists who will be able to engage best with the practitioners are those who can become interested in the challenge of trying to solve their problems, rather than just working to test theories.

Mandel added that he did not see storytelling as a particularly important skill beyond simply having the ability to communicate with the practitioner community in terms that are not full of jargon. "If [researchers] can't talk in a clear way to directors and analysts then they are going to turn those people off," he said, "because they are not going to want to hear about theory X or theory Y or all of these strange terms that psychologists would normally employ when talking with their academic colleagues."

Fein suggested that researchers who are able to work with, hang out with, and gain the trust of those in the practitioner community are likely to be more effective. In particular, researchers should be able to really listen to other people, understand their interests, and try to figure out what they can do that is useful.

Researchers also need to be careful not to oversell what they can do. In particular, practitioners are always interested in getting results they can use as quickly as possible. Researchers need to be honest and objective about just how long it will take to obtain results. They need to be able to say, "I really wish I could help you in the short term, but it would not be fair to you for me to tell you that."

Positive Focus

The last lesson that Meissner offered was the importance of a positive focus. In the eyewitness memory field, he said, the emphasis always seems to be on false memories and mistaken eyewitness identification. Few researchers talk about the positive things that could be done to improve eyewitness identification, and that is a problem. "I think it is really important to have a positive focus," he said. "If you want to change an applied field, you don't go to them wagging your finger saying, 'You are doing this poorly. Stop doing this. Stop doing this.' In fact, what you

need to have are positive alternatives: 'Here is a way that we can improve what you do.'" With that sort of message, researchers are much more likely to listen and respond in a useful way.

IMPLEMENTATION ISSUES

In addition to the general lessons learned from other fields, the workshop participants discussed a number of issues more specific to the task of doing field evaluations of methods from the behavioral sciences applied to intelligence and counterintelligence.

Metrics

One of the issues that was returned to again and again during the workshop was how to judge the effectiveness of various practitioners in the intelligence community. Particularly in the case of analysts, it is difficult to come up with ways to measure outcomes, so a large number of the metrics are based on process instead. Gary McClelland of the University of Colorado reported that of the eight standards listed on an intelligence community directive[2] that he had seen, seven were based on process. Only one of them was based on outcomes: to make accurate judgments and assessments.

This will make it very difficult for researchers to perform useful field evaluations, McClelland said, and it will make it very difficult to convince practitioners to switch to more effective methods. "When we talk about when things will change," he said, "I think it has to come from the intelligence community deciding they will keep score."

Brandon echoed McClelland's comments. Without a clear metric, she noted, it is impossible to set a baseline of where the field is right now, so it is equally impossible to know with any certainly when performance has improved.

McClelland observed that if one looks at the thousands of judgments and forecasts that the intelligence community makes, one would find that most of them are pretty good. But there is absolutely no way of really assessing that, he said, and so the intelligence community ends up being assessed on the basis of a few spectacular events that may be very atypical, such as the failure to foresee the 9/11 attacks and the judgment that Iraq had weapons of mass destruction. But a standardized scoring system would make it possible to keep score. The intelligence community would know how well it was doing and would also be able to see if a

[2]Intelligence Community Directive (ICD) 203—Analytic Standards (June 2007). Available at http://www.fas.org/irp/dni/icd/icd-203.pdf.

new technique improved things or made them worse. And once the intelligence community starts to measure outcomes, then it becomes possible for researchers to compare different methods according to the outcomes that are important to the intelligence community. Researchers should keep in mind that the outcomes they measure should be ones that matter to the intelligence community, rather than ones that seem important to researchers.

Kleinman noted that in all organizations the tendency is to measure what is easy to measure. "It makes for a nice report and a great statistical presentation, but it rarely tells us what we need to know." Things that are really valuable to measure are often quite difficult to measure, he added, and require creativity and constant learning. In the end, he said, it is almost always worth the additional thought and effort.

Take the example of a metric for rating intelligence analyses. "One could argue that good intelligence analysis provides policy makers with meaningful options about what they can do to influence situations. You have just told them they can do these six things, and each has the potential to influence the situation. But how do you measure whether those were good options? You are clear on what you want to try to achieve, but the metric may be incredibly vexing." Still, that doesn't mean it isn't possible to devise suitable metrics, he said. That's what scientists do: find clever ways to measure things. It is often simple, but rarely easy.

Lehner added that having metrics often has the additional value of making problems obvious and creating momentum for change. For example, the DNA exonerations created a very clear metric—people wrongfully convicted on the basis of eyewitness identification—and led to the push to study eyewitness identification with the goal of improving it.

Test and Field Versus Field and Test

Because of the pressure to put new methods out in the field as quickly as possible, one school of thought holds that the best approach is to skip detailed laboratory testing and experimentation and do the testing out in the field once the method has been put to work—the "field-and-test" approach. Others believe that more testing should be done before any method is fielded in order to avoid the problem of practitioners getting attached to—and wasting their time with—methods that eventually prove to be ineffective. Workshop participants discussed the pros and cons of the two approaches.

Meissner commented that there is probably a continuum between the test-and-field and the field-and-test approaches; it is not simply an either/or issue. As a scientist, he tends to be more on the test-and-field side, he said. In part this is because he has found it so difficult to get the

legal system to work with him, so whenever he has had the opportunity, he wanted to make sure he went in with his best stuff. With too many failures, the people he worked with might decide it wasn't worth their while.

Dennis Buede from Innovative Decisions commented that a basic question when trying to determine the correct approach is, How good is good enough? When has something been tested enough to put it into the field? In some domains, more testing is needed early, he said, while others require less. "I would suggest something like APOLLO, which is focused more on thinking, would need a lot less testing prior to fielding than something like a voice stress analyzer where it is conceivable that you may not only be giving the wrong advice but may be sending them in the wrong direction." It is important to apply some common sense when deciding how much testing to do before putting something in the field.

Lehner argued strongly for the field-and-test method. "It is flat out impractical to do full scientific validation before fielding new methods and tools," he said. "The need is urgent, and, quite frankly, good science is just way too slow."

Since it is practically impossible to do the testing first, it will have to be done afterward, and that can be an effective approach if practitioners learn to become effective evaluators of methods. To do this, he said, it is necessary to foster a culture of being open to negative evaluations of current practices—that is, a culture that is just the opposite of the circle-the-wagons mentality that dominates now. Managers and users should be encouraged to ask, "Does this stuff really work? I know it seems to work, but does it really work?" Once a technology or method has been fielded, practitioners should be encouraged to do rigorous evaluations, and negative results should be rewarded. Practitioners should get the message that much of what is fielded may not work, and they need good evaluation practice to sort out what really does work.

By the same token, he said, the scientific community needs to get over the idea that one has to complete all of the scientific research before something is put into the field. What scientists can do to help is to help figure out ways to improve evaluations, to study what constitutes a good process for evaluations based on case experience and personal field experience. Such work will never have the qualities of randomized controlled trials, but it should at least be possible to come up with evaluation methods that are better than what is being done now.

By contrast, Kleinman argued the case for test-and-field. "It has been my experience," he said, "that we would be better off in many cases just not fielding anything new without some high level of confidence—and I mean confidence from the scientific perspective, not the confidence of a program manager." Some people might argue that it is important to try new things

in an effort to find some that work, he said, but when you are talking about national security policy or military affairs, the stakes are way too high. Guessing—or hoping—can often prove to be an expensive proposition with severe strategic consequences.

Eduardo Salas sided with Kleinman. Noting that he has spent the past 25 years conducting field evaluations of systems that attempt to improve human performance in various domains, he said he would never recommend any agency to field and then test. "I think that is dangerous."

Lehner responded by suggesting that the field-and-test approach could lead practitioners to push for better science. Once the practitioners decide that most things don't work and start evaluating everything rigorously, they will quickly get to a point at which they are frustrated with the large number of technologies that fail the evaluations. Ultimately, he suggested, they will say, "Don't send me this stuff until you have good evidence. I already have three things that you have sent out, none of which in the end worked." So it is very possible, he said, that a push toward good science could become a by-product of more aware practitioners.

Getting Practitioners to Use New Techniques

Steven Rieber from the Office of the Director of National Intelligence observed that, depending on the particular area in the intelligence community, it can be easy or difficult to get practitioners to try new techniques. In the area of deception detection, people tend to want tools immediately. But intelligence analysts are often reluctant to use new tools or techniques. So he asked the group if there was any research or anecdotal evidence to suggest how best to convince these practitioners to try new techniques.

Mandel responded that time constraints are one of the biggest issues for analysts. During training, he said, the analysts are taught various methods, such as analysis of competing hypotheses, but when the analysts get on the job, "they say they don't have the time to use those things because they just get bogged down right away and then [are] always trying to catch up." He added that he believes that the organizational constraints that affect the uptake of even good techniques are an important topic for research.

Jim Powlen of Logos Technologies expanded on Mandel's comments. In his discussions with analysts, he said, he finds them as eager as anybody for more effective tools, but at the same time they complain that they have too many of them. They say, "I have 500 tools. I have more tools then I can possibly remember or ever use. I don't need another tool."

But if you pursue it a little further, he said, you discover that what they really want is one-stop shopping—a suite that will help them con-

solidate the information that they need, so that instead of spending 80 percent of their time gathering information and 20 percent doing analysis, they can reverse it to spend only 20 percent of their time bringing in relevant information and 80 percent on analysis. "My perception is that they're as eager for help in the technology arena as anybody else," he said. "They have too many single-action tools, and that isn't really helping them."

Whitehurst offered his perspective from the education field: "A romantic idea that I used to hold is that if you found out something that was truly useful to practice, and you made it available in a pamphlet or a publication, and you even got practitioners to read it, that they would change their behavior as a result. It was, as I label it in retrospect, a hopelessly romantic view." He now believes that, in many cases at least, the uptake of new technologies will not happen unless there are contingencies that require it to happen. "Nothing changes unless there are contingencies in the system to require change, accountability, or on-the-job requirements, or something. Then the teachers will change just like police change, just like university professors change, just like intelligence officers change—because they have to."

Intelligence Institute

Several workshop speakers and participants spoke of the value of creating an intelligence institute dedicated to producing solid research on issues of importance to intelligence, much as the National Institutes of Health produce solid research on issues of importance to health. There were a number of arguments for such an institute.

Thomason offered two basic reasons for creating an intelligence institute. The first is that there really isn't an internal research tradition within the intelligence community, and an intelligence institute could go a long way toward establishing such an internal tradition. The second is that there are many well-trained people outside the intelligence community who would be very interested in working on intelligence-related issues if the opportunity arose, and an intelligence institute could, if it was well financed, accelerate the collaboration process.

Robert Boruch commented that unless a clear place for scientific evidence is set aside in a governmental organization, no science will be introduced into that organization. That is the idea behind the National Science Foundation, for example. Furthermore, once a science-based entity is set up, it is important to protect it and its science from nonscientific influences. For instance, federal statistical agencies such as the Census Bureau and the Bureau of Labor Statistics have special statutory provisions intended to insulate them from the influences of theology, politics, ideology, and so

on. Understanding how to build that protection into an intelligence institute is very important, he said.

In many ways, the Defense Personnel Security Research Center, or PERSEREC, parallels what people in the workshop were discussing as an intelligence research institute, and Eric Lang of PERSEREC described a bit of its history to offer some insight into what it might take to set up an intelligence institute.

After a rash of espionage cases in the mid-1980s, a security review commission recommended that the Department of Defense develop an organic research capability to understand the problems better. PERSEREC was set up with a sunset clause: it had three years to prove its worth, or it would be shut. "What we did," Lang said, "is develop a strategic plan that had a mix of quick-hitting research studies and longer term programmatic research, and we became the institutional memory for DoD and for much of the rest of the government because there is no other similar size research entity dedicated to personnel security."

There was constant pressure to provide devices and methods that could be used immediately—to take the "low-hanging fruit"—and PERSEREC did provide some of this. "This is part of how we earn our keep," Lang said. But PERSEREC also devotes a significant portion of its time to long-term programmatic research, and that has paid off. Even though some of the studies have taken three years, five years, or longer, they are valued and many have resulted in policy improvements at the DoD and national levels. The clients at the undersecretary level value the programmatic research, Lang said. "We have a critical mass of mostly Ph.D.-level social scientists and psychologists who provide a stable source of knowledge and hands-on experience for understanding personnel security needs, working with the key players in the field and leadership positions, and conducting both long-term and short-term research. And we can make a case for the practical value that both kinds of research provide."

Lang argued that the intelligence community needs something similar—an organic, ongoing research infrastructure and capability, rather than just commissioning an isolated project here and a collaboration there. Part of the value of PERSEREC, he said, is that it has been around for more than 20 years. "People in the community know our staff, track record, and capabilities. They know we will help them think through the problem, do the research, and, if needed, help with implementation and follow-up evaluation. But it takes that kind of ongoing institutional memory and critical mass of applied and basic researchers to get that job done."

Either the Defense Intelligence Agency or the Office of the Director of National Intelligence is a logical place for an intelligence research institute, Lang said. But regardless of location, it is important for it to be established with the proper charter, one that sets up a suitable research

capability that allows the institute to delve into issues on a regular basis and not simply at workshops or in the form of consensus studies.

None of this is easy. "This is a very tough problem," Fein commented. The workshop discussions, particularly those that presented experiences from other fields, made it clear there are many obstacles to effective field evaluations of behavioral science techniques. They also made it clear that such evaluations are possible with the right approach and enough effort—and, furthermore, that such evaluations are indeed crucial to determining which methods should be put to work. It requires patience and a long-term view, but it can be done. "I emerge from these discussions sobered but actually more hopeful than before," Fein said, if only because the workshop demonstrated that quite a number of good minds are already at work on the problem.

References

Bhatt, S., and Brandon, S.E. (2008). *Review of the preliminary credibility assessment screening system (PCASS).* Unpublished manuscript, Washington, DC.

Bhatt, S., and Brandon, S.E. (2009). *Review of voice stress-based technologies for the detection of deception.* Unpublished manuscript, Washington, DC.

Brander, G. (2009). *A U.K. perspective.* Presentation at the Workshop on Field Evaluation of Behavioral and Cognitive Sciences-Based Methods and Tools for Intelligence and Counterintelligence, September 22-23, National Academies, Washington, DC. Available: http://nationalacademies.org/bbcss/Field_Evaluation_Workshop_Presentations.html [accessed February 2010].

Connors, E., Lundregan, T., Miller, N., and McEwen, T. (1996). *Convicted by juries, exonerated by science.* Washington, DC: U.S. Department of Justice, Office of Justice Programs, National Institute of Justice.

Fisk, C. (1972). The Sino-Soviet border dispute: A comparison of the conventional and Bayesian methods for intelligence warning. *Studies in Intelligence, 16*(2), 53-62.

Grove, W.M., Zald, D.H., Hallberg, A.M., Lebow, B., Snitz, E., and Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12,* 19-30.

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

Innocence Project. (2009a). *Innocence Project case profiles.* Available: http://www.innocence project.org/know/ [accessed January 2010].

Innocence Project. (2009b). *The causes of wrongful conviction*. Available: http://www.innocence project.org/understand/ [accessed September 2009].

Kirkpatrick, D.L. (1994). *Evaluating training programs: The four levels.* San Francisco: Berrett-Koehler.

Lindsay, R.C.L., and Wells, G.L. (1985). Improving eyewitness identification from lineups: Simultaneous versus sequential lineup presentations. *Journal of Applied Psychology, 70*, 556-564.

Lum, C., Kennedy, L.W., and Sherley, A.J. (2006). The effectiveness of counter-terrorism strategies: A Campbell systematic review. *Journal of Experimental Criminology*, *2*(4), 489-516.

Lum, C., Koper, C., and Telep, C.W. (2009). Evidence-based policing matrix. Available: http://gemini.gmu.edu/cebcp/matrix.html [accessed January 2010].

Meissner, C.A. (2009). *Eyewitness (mis)identification: How errors of memory can lead to wrongful conviction*. Presented at the Actual Innocence Conference, Plano, TX.

Munsterberg, H. (1908). *On the witness stand: Essays on psychology and crime.* New York: Doubleday, Page.

National Research Council. (1999). *Improving student learning: A strategic plan for education research and its utilization.* Committee on a Feasibility Study for a Strategic Education Research Program, Commission on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press.

National Research Council. (2003). *The polygraph and lie detection.* Committee to Review the Scientific Evidence on the Polygraph. Board on Behavioral, Cognitive, and Sensory Sciences and Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2008). *Emerging cognitive neuroscience and related technologies.* Committee on Military and Intelligence Methodology for Emergent Neurophysiological and Cognitive/Neural Science Research in the Next Two Decades. Standing Committee for Technology Insight—Gauge, Evaluate, and Review Division on Engineering and Physical Sciences. Board on Behavioral, Cognitive, and Sensory Sciences, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

National Research Council. (2009). *Opportunities in neuroscience for future army applications.* Committee on Opportunities in Neuroscience for Future Army Applications. Board on Army Science and Technology, Division on Engineering and Physical Sciences. Washington, DC: The National Academies Press.

National Research Council. (2010). *Strengthening scientific research and development at the National Institute of Justice.* Committee on Assessing the Research Program of the National Institute of Justice. Center for Economic, Governance, and International Studies, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Ruger, T., Kim, P., Martin, A., and Quinn, K. (2004). The Supreme Court Forecasting Project: Legal and political science approaches to Supreme Court decision making. *Columbia Law Review, 104*, 1150. Available: http://www.law.upenn.edu/cf/faculty/truger/workingpapers/104ColumLR1150.pdf [accessed January 2010].

Sherman, L.W. (1998). *Evidence-based policing*. Washington, DC: Police Foundation.

Sticha, P., Buede, D., and Rees, R.L. (2005). *APOLLO: An analytical tool for predicting a subject's decision making*. Presented at the International Conference on Intelligence Analysis Methods and Tools. May 2-6, McLean, VA.

Technical Working Group for Eyewitness Evidence. (1999). *Eyewitness evidence: A guide for law enforcement.* Rockville, MD: U.S. Department of Justice, National Institute of Justice.

Technical Working Group for Eyewitness Evidence. (2003). *Eyewitness evidence: A trainer's manual for law enforcement.* Rockville, MD: U.S. Department of Justice, National Institute of Justice.

U.S. Department of Education. (2007). *Report of the Academic Competitiveness Council.* Washington, DC.

Wells, G.L., Small, M., Penrod, S., Malpass, R.S., Fulero, S.M., and Brimacombe, C.A.E. (1998). Eyewitness identification procedures: Recommendations for lineups and photospreads. *Law and Human Behavior, 22*(6), 603-647.

Appendix A

Workshop Agenda and Participants

AGENDA

Workshop on Field Evaluation of Behavioral and Cognitive Sciences-Based Methods and Tools for Intelligence and Counterintelligence

Tuesday, September 22, 2009

8:00 a.m.	Workshop Check-In
8:20	Welcome *Barbara Wanchisen,* director, Board on Behavioral, Cognitive, and Sensory Sciences, National Research Council
8:30	Background and Committee Genesis *Philip Rubin,* committee chair and CEO, Haskins Laboratories

EXAMPLES OF BEHAVIORAL TOOLS AND METHODS

Moderator: Robert Fein, forensic psychologist, Harvard Medical School and committee member

8:45	Preliminary Credibility Assessment Screening System (PCASS) *Donald Krapohl*, special assistant to the academy director, Defense Academy for Credibility Assessment

9:15	Voice Stress Technologies *Philip Rubin*
9:45	Break
10:00	APOLLO *Charles Twardy*, research assistant professor, George Mason University
10:30	Alternative Competing Hypotheses (ACH) *Neil Thomason*, senior lecturer (retired), University of Melbourne and committee member
11:00	Response and Discussion Respondent Reactions *Steven Kleinman*, consultant and strategist, Intelligence and National Security Policy *Paul Lehner*, chief engineer, Information Technology Division, Center for Integrated Intelligence Systems, MITRE Corporation General Discussion
12:30 p.m.	Working Lunch

MODELS OF EVALUATION IN OTHER FIELDS
Moderator: Philip Rubin

1:30	Challenges in Field Evaluation for Education *Grover (Russ) Whitehurst,* senior fellow, Governance Studies, and director, Brown Center on Education Policy
2:00	Validation and Evaluation Practices in the Health Sciences *Lisa J. Colpe,* senior scientist, Division of Services and Intervention Research, National Institute of Mental Health
2:30	Evidence-Based Practices in Criminal Justice *Cynthia Lum,* deputy director, Center for Evidence-Based Crime Policy and assistant professor, Administration of Justice Department, George Mason University
3:00	Human Factors and Organizational Psychology: Application to Training Evaluation *Eduardo Salas,* professor, Institute for Simulation and Training, University of Central Florida and committee member
3:30	Break

3:45	Response and Discussion Respondent Reaction: Possible Application to Intelligence and Counterintelligence Community Needs *Robert Boruch,* professor, Graduate School of Education and Statistics, University of Pennsylvania and committee member General Discussion
4:55	Closing Comments *Philip Rubin*
5:00	Adjourn

Wednesday, September 23, 2009

9:00 a.m.	Workshop Check-In
9:10	Welcome *Barbara Wanchisen*
9:20	Background and Review of Day One *Philip Rubin*

ETHICAL, REGULATORY, AND CULTURAL CONSIDERATIONS
Moderator: Philip Rubin

9:30	Ethical Challenges in Translating Psychophysiology and Neuroscience to Technology for Intelligence and Counterintelligence *Adil Shamoo*, editor-in-chief, *Accountability in Research,* University of Maryland School of Medicine Commentary *Jonathan Moreno*, David and Lyn Silfen University Professor, Department of Medical Ethics and Department of History and Sociology of Science, University of Pennsylvania and committee member Discussion
10:30	Break

INTERNATIONAL PERSPECTIVES AND APPLICATION OF LESSONS FROM POLICING
Moderator: Philip Rubin

10:45	A UK Perspective *George Brander*, Ministry of Defence, United Kingdom
11:15	Canadian Defense Validation Efforts

David Mandel, defence scientist and group leader, Defence Research and Development Canada, Toronto

11:45 Application of Lessons from Forensics
Christian Meissner, associate professor and chair, Institutional Review Board, Department of Psychology, University of Texas at El Paso

12:15 p.m. Working Lunch

1:15 Response and Discussion
Committee Reaction
General Discussion

SUMMATION AND IMPLICATIONS

Moderator: Philip Rubin

2:45 Summation, Discussion, and Suggestions for Future Evaluations
Summative Comments
Robert Fein
Committee Reactions
General Discussion

3:45 Closing Comments
Philip Rubin

4:00 Adjourn

PARTICIPANTS

Planning Committee Members:

Robert F. Boruch, University of Pennsylvania
Robert A. Fein, Harvard Medical School
Jonathan D. Moreno, University of Pennsylvania
Philip E. Rubin (*Chair*), Haskins Laboratories
Eduardo Salas, University of Central Florida
Neil Thomason, University of Melbourne
Carol H. Weiss, Harvard University

Other Workshop Presenters:

George Brander, Defence Science and Technology Laboratory, Cyber & Influence Centre; UK Ministry of Defence Human Factors Team
Lisa Colpe, National Institute of Mental Health
Steven M. Kleinman, consultant and strategist, Intelligence and National Security Policy
Donald Krapohl, Defense Academy for Credibility Assessment

Paul Lehner, MITRE Corporation
Cynthia Lum, Center for Evidence-Based Crime Policy, George Mason University
David R. Mandel, Defence Research and Development Canada
Christian Meissner, Department of Psychology, University of Texas at El Paso
Adil E. Shamoo, School of Medicine, University of Maryland
Charles R. Twardy, George Mason University
Grover "Russ" Whitehurst, Brown Center on Education Policy

National Research Council Staff:
Laudan Aron, Division of Behavioral and Social Sciences and Education
Dan Melnick, National Research Council
Patricia Morison, Center for Education
Mary Ellen O'Connell, Board on Behavioral, Cognitive, and Sensory Sciences
Robert Pool, consultant
Miron Straf, Division of Behavioral and Social Sciences and Education
Barbara Wanchisen, Board on Behavioral, Cognitive, and Sensory Sciences
Renée L. Wilson Gaines, Board on Behavioral, Cognitive, and Sensory Sciences

Registered Attendees:
Hal Arkes, Ohio State University
Zunair Ashfaq, University of Pennsylvania
Emma Barrett, UK Ministry of Defence
Monique E. Beaudoin, National Defense University, Ottawa
Ronald Benefield, Office of the Naval Counterintelligence Executive
Sujeeta Bhatt, Defense Intelligence Agency
Anthony Boemio, Booz Allen Hamilton
Susan Brandon, Defense Counterintelligence and Human Intelligence Center, Defense Intelligence Agency
Troy Brown, Defense Academy for Credibility Assessment
James Bruce, RAND Corporation
Dennis Buede, Innovative Decisions, Inc.
Michael Cassidy, Marymount University
Ron Ceasar, Ron Ceasar Photography
Paul Chatelier, Naval Postgraduate School
Michael Cheek, ExecutiveBiz
A. Egon Cholakian, Interdisciplinary Research Development Fund Project
Brian Colder, National Geospatial-Intelligence Agency

Keith Devereaus, U.S. Department of Homeland Security-Citizenship and Immigration Services
Ivy Estabrooke, Office of Naval Research
Pamela Flattau, Institute for Defense Analyses Science and Technology Policy Institute
Angelyn Flowers, University of the District of Columbia
Melanie Goodrich, Office of Naval Research
Leslie Goodyear, National Science Foundation
Julie Gravallese, MITRE Corporation
Hal Greenwald, MITRE Corporation
Joe Heaps, National Institute of Justice
Kristin Heckman, MITRE Corporation
Cheryl Hendrickson Caster, American Institutes for Research
Georgia Holmer, Pherson Associates, LLC
Alexis Jeannotte, Avian Engineering
David Kamien, Mind-Alliance Systems, LLC
Eric Kaufman, National Counterterrorism Center
Heather Kelly, American Psychological Association
A.T. Kendall, social research consultant
Mary Lee Kingsley, M.L. Kingsley, LLC
Adam Korobow, Logistics Management Institute
Robert Krikorian, U.S. Department of State
Howard Kurtzman, American Psychological Association
Thomas LaHann, Science Applications International Corporation
Eric Lang, Defense Personnel Security Research Center
Thomas Lawrence, Pennsylvania Army National Guard
Richard Lempert, Department of Homeland Security Science and Technology
Heather Leon, Navy Marine Corps Intelligence Training Center
Tod Levitt, George Mason University
Tiffany Lightbourn, Department of Homeland Security–U.S. Citizenship and Immigration Services
Deborah Loftis, Defense Intelligence Agency
Martha Lorber, MITRE Corporation
Kel McClanahan, National Security Counselors
Gary McClelland, University of Colorado
Robert McCreight, George Washington University
Nancy Merritt, National Institute of Justice
Erica Michael, University of Maryland Center for Advanced Study of Language
Avra Michelson, MITRE Corporation
Thomas Moore, Defense Advantage Research Projects Agency
Dwayne Norris, American Institutes for Research

William Norris, Defense Academy for Credibility Assessment
J. O'Connor, Department of Homeland Security Human Factors/ Behavioral Sciences Division
Jason Ogden, Customs and Border Protection
Randolph Pherson, Pherson Associates, LLC
Amy Pollick, Association for Psychological Science
Matthew Potts, Marine Corps Intelligence Association, Inc.
Jim Powlen, Logos Technologies, Inc.
Joseph Psotka, U.S. Army Research Institute
Leslie Richards, University of the District of Columbia
Steven Rieber, Office of the Director of National Intelligence
Michael Rugnetta, Center for American Progress
Adam Russell, Intelligence Advanced Research Projects Activity
Christina Saylor, U.S. Special Operations Command
Grace Scarborough, Pherson Associates, LLC
Dylan Schmorrow, Office of the Secretary of Defense
Rachael Scholz, Booz Allen Hamilton
Nathan Schwade, Sandia National Laboratories
Alan Schwartz, PolicyFutures, LLC
Dan Schwartz, Schwartz and Schwartz, LLC
Allison Smith, Department of Homeland Security Science and Technology Directorate
David L. Smith, Navy Marine Corps Intelligence Training Center
Robert A. Smith, University of Maryland
Jonathan Snider, Defense Threat Reduction Agency
Kevin Spence, Department of Homeland Security
Barry Spodak, Action Training Institute
Frank Stech, MITRE Corporation
Scott Tousley, MITRE Corporation
Edwin Urie, Henley-Putnam University
Anthony B. Veney, U.S. Department of Defense
Jeremy Wacksman, U.S. Department of the Navy, U.S. Department of Defense
Mark Weiss, National Science Foundation
Joe Wholey, University of Southern California
Deborah York, Phoenix Consulting Group
Laura Zimmerman, Applied Research Associates

Appendix B

Relevant Readings

INTELLIGENCE AND INTELLIGENCE ANALYSIS-GENERAL

Bennett, M., and Waltz, E. (2007). *Counterdeception: Principles and applications for national security.* Boston: Artech House.

Bruce, J.B. (2008). Making analysis more reliable: Why epistemology matters to intelligence In R.Z. George and J.B. Bruce (Eds.), *Analyzing intelligence: Origins, obstacles, and innovations* (pp. 171-190). Washington, DC: Georgetown University Press.

Bruce, J.B., and Bennett, M. (2008). Foreign denial and deception: Analytical imperatives. In R.Z. George and J.B. Bruce (Eds.), *Analyzing intelligence: Origins, obstacles, and innovations* (pp. 122-137). Washington, DC: Georgetown University Press.

Clark, R.M. (2007). *Intelligence analysis: A target-centric approach.* Washington, DC: CQ Press.

Davis, J. (1999). Introduction: Improving intelligence analysis at CIA: Dick Heuer's contribution to intelligence analysis. In R.J. Heuer, Jr. (Ed.), *Psychology of intelligence analysis* (pp. xiii-xxv). Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency.

Folker, R. (2000). *Intelligence analysis in theater joint intelligence centers: An experiment in applying structured methods.* Occasional Paper 7. Washington, DC: Joint Military Intelligence College. Available: http://www.au.af.mil/au/awc/awcgate/dia/analysis_structured.pdf [accessed February 2010].

George, R.Z., and Bruce, J.B. (2008). *Analyzing intelligence: Origins, obstacles, and innovations.* Washington, DC: Georgetown University Press.

Goodson, R., and Wirtz, J. (2008). Strategic denial and deception: The twenty-first century challenge. In R. Goodson and J.J. Wirtz (Eds.), *Strategic denial and deception: The twenty-first century challenge* (pp. 1-14). London: Transaction.

Heuer, R.J., Jr. (Ed.). (1979). *Quantitative approaches to political intelligence: The CIA experience.* Boulder, CO: Westview Press.

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

Johnston, R. (2005). *Analytic culture in the U.S. intelligence community*. Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency. Available: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/analytic-culture-in-the-u-s-intelligence-community/analytic_culture_report.pdf [accessed February 2010].

Kennedy, R. (2008). *Of knowledge and power: The complexities of national intelligence.* Wesport, CT: Praeger Security International.

Kent, S. (1949). *Strategic intelligence for American world policy*. Princeton, NJ: Princeton University Press.

Kent, S. (1955). The need for an intelligence literature. Reprinted in D.P. Steury (Ed.), *Sherman Kent and the Board of National Estimates*. Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency. Available: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/2need.html [accessed February 2010].

Kent, S. (1964). Words of estimative probability. Reprinted in D.P. Steury (Ed.), *Sherman Kent and the Board of National Estimates*. Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency. Available: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/6words.html [accessed February 2010].

Lehner, P. (2009). *The objective analysis of analysis*. Paper presented at the Community of Interest for the Practice and Organization of Intelligence Ottawa Roundtable "What Can the Cognitive and Behavioral Sciences Contribute to Intelligence Analysis?: Towards a Collaborative Agenda for the Future," February, Meech Lake, Quebec.

Mandel, D.R. (2009). *Setting the stage: The role of science in applied communities*. Paper presented at the Community of Interest for the Practice and Organization of Intelligence Ottawa Roundtable "What Can the Cognitive and Behavioral Sciences Contribute to Intelligence Analysis?: Towards a Collaborative Agenda for the Future," February, Meech Lake, Quebec.

Marrin, S. (2009). Training and educating U.S. intelligence analysts. *International Journal of Intelligence and Counter-Intelligence, 22,* 131-146.

Milgram, S. (1974). *Obedience to authority: An experimental view.* New York: Harper & Row.

Moore, D.T. (2006). *Critical thinking and intelligence analysis*. Occasional Paper Number 14. Joint Military Intelligence College, May.

Rieber, S. (2004). Intelligence analysis and judgmental calibration. *International Journal of Intelligence and Counter-Intelligence, 17,* 97-112.

Rieber, S., and Thomason, N. (2006). Toward improving intelligence analysis: Creation of a National Institute of Analytic Methods. *Studies in Intelligence, 49*(4), 71.

Sims, J.E., and Gerber, B. (2005). *Transforming U.S. intelligence.* Washington, DC: Georgetown University Press.

Swenson, R.G. (2002). Meeting the community's continuing need for an intelligence literature. *Defense Intelligence Journal, 11*(2), 87-98.

Warner, M. (2009). Sources and methods for the study of intelligence. In L.K. Johnson (Ed.), *Handbook of intelligence studies* (pp. 17-27). New York: Routledge.

Weinstein, N. (1980). Unrealistic optimism about future life events. *Journal of Personality and Social Psychology*, *39*, 806-820.

Westerfield, H.B. (Ed.). (1995). *Inside CIA's private world: Declassified articles from the agency's internal journal, 1955-1992*. New Haven: Yale University Press.

Wirtz, J.J. (2009). The American approach to intelligence studies. In L.K. Johnson (Ed.), *Handbook of intelligence studies* (pp. 28-38). New York: Routledge.

Yarger, H.R. (2008). *Strategy and the national security professional: Strategic thinking and strategic formulation in the 21st century.* Westport, CT: Praeger.

ALTERNATIVE COMPETING HYPOTHESES

Billman, D., Convertino, G., Shrager, J., Massar, J.P., and Pirolli, P. (2006). *Collaborative intelligence analysis with CACHE and its effects on information gathering and cognitive bias*. Technical Report. Palo Alto, CA: Palo Alto Research Center.

Cheikes, B.A., Brown, M.J., Lehner, P.E., and Adelman, L. (2004). *Confirmation bias in complex analyses.* Technical Report. Bedford, MA: MITRE Center for Integrated Intelligence Systems.

Cluxton, D., and Eick, S.G. (2005). *DECIDETM: Hypothesis visualization tool.* Presented at the International Conference on Intelligence Analysis Methods and Tools, May 2-6, McLean, VA.

Convertino, G., Billman, D., Pirolli, P., Massar, J.P., and Shrager, J. (2006). *Collaborative intelligence analysis with CACHE: Bias reduction and information coverage*. Technical Report. Palo Alto, CA: Palo Alto Research Center.

Convertino, G., Billman, D.O., Pirolli, P.L., Massar, J.P., and Shrager, J. (2008). The CACHE study: Group effects in computer-supported collaborative analysis. *Computer Supported Cooperative Work, 17*, 353-393.

Good, L., Shrager, J., Stefik, M., Pirolli, P., and Card, S. (2004). *ACH0: A tool for analyzing competing hypotheses.* Technical description for Version 1.0. Palo Alto, CA: Palo Alto Research Center.

Heuer, R.J., Jr. (2008). Computer-aided analysis of competing hypotheses. In R.Z. George and J.B. Bruce (Eds.), *Analyzing intelligence: Origins, obstacles, and innovations* (pp. 251-265). Washington, DC: Georgetown University Press.

Heuer, R.J., Jr. (2008). *Improving intelligence analysis with ACH*. Available: http://www.pherson.org/PDFFiles/Heuer-ImprovingIntelligenceAnalysiswithACH.pdf [accessed February 2010].

Lehner, P.E., Adelman, L., Cheikes, B.A., and Brown, M.J. (2008). Confirmation bias in complex analyses. *IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans, 38*(3), 584-592.

Pherson, R. (2008). *Collaborative analysis of competing hypotheses (C-ACH).* Available: http://www.pherson.org/pdffiles/Collaborative-ACH.pdf [accessed February 2010].

Pirolli, P. (2006). *Assisting people to become independent learners in the analysis of intelligence*. Technical Report. Palo Alto, CA: Palo Alto Research Center.

Pope, S., and Jøsang, A. (2005). *Analysis of competing hypotheses using subjective logic*. Tenth International Command and Control Research and Technology Symposium, June, McLean, VA.

Stech, F.J., and Elsaesser, C. (2005). *Deception detection by analysis of competing hypotheses*. Technical Report. McLean, VA: MITRE Corporation.

Valtorta, M., Dang, J., Goradia, H., Huang, J., and Huhns, M. (2005). *Extending Heuer's analysis of competing hypotheses method to support complex decision analysis*.Presented at the International Conference on Intelligence Analysis Methods and Tools, May 2-6, McLean, VA.

Wheaton, K.J., and Chido, D.E. (2006). Structured analysis of competing hypotheses: Improving a tested intelligence methodology. *Competitive Intelligence Magazine, 9*(6), 12-15.

APPLIED BAYESIAN ANALYSIS

Dawes, R.M., and Corrigan. B. (1974). Linear models in decision making. *Psychological Bulletin, 81*, 95-106.

Dawes, R., Faust, R., and Meehl, P. (1989). Clinical versus actuarial judgment. *Science, 243*(4899), 1668-1674.

Digman, J.M. (1990). Personality structure: An emergence of the five-factor model. *Annual Review of Psychology, 41*, 417-440.

Fisk, C. (1972). The Sino-Soviet border dispute: A comparison of the conventional and Bayesian methods for intelligence warning. *Studies in Intelligence, 16*(2), 53-62.

Grove, W.M., Zald, D.H., Lebow, B.S., Snitz, B.E., and Nelson, C. (2000). Clinical versus mechanical prediction: A meta-analysis. *Psychological Assessment, 12*(1), 19-30. Available: http://www.psych.umn.edu/faculty/grove/096clinicalversusmechanicalprediction.pdf [accessed February 2010].

Kahneman, D., Slovic, P., and Tversky, A. (Eds.). (1982). *Judgment under uncertainty: Heuristics and biases.* Cambridge, UK: Cambridge University Press.

Kent, S. (1964). Words of estimative probability. Reprinted in D.P. Steury (Ed.), *Sherman Kent and the Board of National Estimates*. Washington, DC: Center for the Study of Intelligence, Central Intelligence Agency. Available: https://www.cia.gov/library/center-for-the-study-of-intelligence/csi-publications/books-and-monographs/sherman-kent-and-the-board-of-national-estimates-collected-essays/6words.html [accessed February 2010].

Neapolitan, R.E. (1990). *Probabilistic reasoning in expert systems*. New York: Wiley.

Nisbett, R.E., and Ross, L. (1980). *Human inference: Strategies and shortcomings of social judgment.* Englewood Cliffs, NJ: Prentice-Hall.

Pearl, J. (1988). *Probabilistic reasoning in intelligent systems*. San Mateo, CA: Morgan Kaufmann.

Schum, D.A. (1994). *Evidential foundations of probabilistic reasoning*. New York: Wiley.

Sticha, P., Buede, D., and Rees, R.L. (2005). *APOLLO: An analytical tool for predicting a subject's decision making.* Presented at the International Conference on Intelligence Analysis Methods and Tools, May 2-6, McLean, VA.

Wallsten, T.S., Budescu, D.V., and Zwick, R. (1993). Comparing the calibration and coherence of numerical and verbal probability judgments. *Management Science, 39,* 176-190.

PRELIMINARY CREDIBILITY ASSESSMENT SCREENING SYSTEM (PCASS)

Battelle Memorial Institute. (2007). *Efficacy of prototype credibility assessment technology: PCASS final report.* Technical Report. Columbus, OH: Battelle Memorial Institute (Defense Academy for Credibility Assessment).

Ben-Shakhar, G., and Eitan, E. (2003). The validity of psychophysiological detection of information with the guilty knowledge test: A meta-analytic review. *Journal of Applied Psychology, 88*(1), 131-151.

Bhatt, S., and Brandon, S.E. (2008). *Review of the preliminary credibility assessment screening system (PCASS).* Unpublished manuscript, Washington, DC.

Clapper, J.R., Jr. (2007). *Operational approval of the Preliminary Credibility Assessment Screening System (PCASS).* Washington, DC: U.S. Department of Defense.

Dedman, B. (2008a). *New U.S. weapon: Hand-held lie detector.* Available: http://www.msnbc.msn.com/id/23926278/ [accessed February 2010].

Dedman, B. (2008b). *What is the PCASS and how does it work?* Available: http://www.msnbc.msn.com/id/24015982/ [accessed February 2010].

Harris, J.C., and McQuarrie, A.D. (2006). *The Preliminary Credibility Assessment System embedded algorithm description and validation results*. Technical Report. Laurel, MD: Applied Physics Laboratory, Johns Hopkins University (Counterintelligence Field Activity).

National Research Council. (2003). *The polygraph and lie detection.* Committee to Review the Scientific Evidence on the Polygraph. Board on Behavioral, Cognitive, and Sensory Sciences and Committee on National Statistics, Division of Behavioral and Social Sciences and Education. Washington, DC: The National Academies Press.

Senter, S., Waller, J., and Krapohl, D. (2006). *Validation studies for the Preliminary Credibility Assessment Screening System (PCASS).* Technical Report. Fort Jackson, SC: U.S. Department of Defense Polygraph Institute.

Thompson, D. (2007). *Evaluation of the Preliminary Credibility Assessment Screening System (PCASS).* Camp Cropper, Iraq, Department of the Army. Available: http://antipolygraph.org/documents/PCASS_Camp_Cropper_memo.pdf [accessed February 2010].

VOICE STRESS ANALYSIS

Baker Group. (2007). *What is DVSA*. Available: http://bakerdvsa.com/whatisdvsa.htm [accessed February 2010].

Bhatt, S., and Brandon, S.E. (2009). *Review of voice stress–based technologies for the detection of deception.* Unpublished manuscript, Washington, DC.

Brown, T.E., Senter, S.M., and Ryan, A.H., Jr. (2003). *Ability of the Vericator™ to detect smugglers at a mock security checkpoint.* Report No. DoDPI00-P-0024. Fort Jackson, SC: U.S. Department of Defense Polygraph Institute.

Cestaro, V.L. (1995). *A comparison between decision accuracy rates obtained using the polygraph instrument and the computer voice stress analyzer (CVSA) in the absence of jeopardy*. Report No. DoDPI95-R-0002. Fort McClellan, AL: U.S. Department of Defense Polygraph Institute.

Cestaro, V.L. (1996). *A comparison of accuracy rates between detection of deception examinations using the polygraph and the computer voice stress analyzer in a mock crime scenario*. Report No. DoDPI-R-0004. Fort McClellan, AL: U.S. Department of Defense Polygraph Institute.

Diogenes Company. (2008). *Diogenes Digital Voice Stress Analysis™ DDVSA™ features.* Kissimmee, FL: Author.

Haddad, D., Walter, S., Ratley, R., and Smith, M. (2001). *Investigation and evaluation of voice stress analysis technology*. In-house technical memorandum AFRL-IF-RS-TM-2001-7. Rome, NY: Air Force Research Laboratory Information Directorate.

Harnsberger, J. D., Hollien, H., Martin, C.A., and Hollien, K.A. (2009). Stress and deception in speech: Evaluating layered voice analysis. *Journal of Forensic Science, 54*(3), 642-650.

Hollien, H., and Harnsberger, J.D. (2006). *Voice stress analyzer instrumentation evaluation*. Final Report CIFA Contract FA 4814-04-0011. Gainesville, FL: Institute for Advanced Study of the Communication Processes, University of Florida.

Hollien, H., Harnsberger, J.D., Martin, C.A., and Hollien, K.A. (2008). Evaluation of the NITV CVSA. *Journal of Forensic Science, 53*(1), 183-193.

Holman, D. (2005). Nothing but the truth. *The American Spectator*. Available: http://spectator.org/archives/2005/12/15/nothing-but-the-truth [accessed February 2010].

Hopkins, C.S., Ratley, R.J., Benincasa, D.S., and Greico, J.J. (2005). *Evaluation of voice stress analysis technology*. 38th Hawaii International Conference on System Sciences, IEEE.

Janniro, M.J., and Cestaro, V.L. (1996). *Effectiveness of detection of deception examinations using the computer voice stress analyzer*. Report No. DoDPI-R-0005. Fort McClellan, AL: U.S. Department of Defense Polygraph Institute.

Krapohl, D., Ryan, A.H., and Shull, K.W. (2002). Voice stress devices and the detection of lies. *Policy Review*. Available: http://hnspolygraph.com/media/voice_stress_devices_and_the_detection_of_lies.pdf [accessed February 2010].

Meyerhoff, J.L., Saviolakis, G.A., Koenig, M.L., and Yourick, D.L. (2000). *Physiological and biological measures of stress compared to voice stress analysis using the computer voice stress analyzer (CVSA).* Report No. DoDPI98-R-0004. Fort Jackson, SC: U.S. Department of Defense Polygraph Institute.

Nemesysco. (2005). *Old versions—Vericator™*. Available: http://www.nemesysco.com/Prod_VERICATOR.html [accessed February 2010].

Nemesysco. (2008). *LVA—Layered voice analysis—Nemesysco's technologies*. Available: http://www.nemesysco.com/technology-lvavoiceanalysis.html [accessed February 2010].

NITV. (2007). *Structured interview assessment of the field use of the voice stress analyzer technology*. Available: http://www.cvsa1.com/VSAAssessment.pdf [accessed February 2010].

NITV. (2008a). *Agencies utilizing the CVSA*. Available: http://www.cvsa1.com/Agencies using.htm [accessed February 2010].

NITV. (2008b). *U.S. special operations command independent evaluation validates the CVSA.* Available: http://www.cvsa1.com/USSpecialOper.htm [accessed February 2010].

Palmatier, J.J. (1999). The computerized voice stress analyzer: Modern technological innovation or "the emperor's new clothes"? *ABA General Practice, Solo & Small Firm Division Magazine*. Available: http://www.abanet.org/genpractice/magazine/1999/jun/palmatr.html [accessed February 2010].

Palmatier, J.J. (2005). Assessing credibility: ADVA technology, voice and voice stress analysis. In R.J. Montgomery and W.J. Majeski (Eds.), *Corporate Investigations* (pp. 37-60). Tuscon, AZ: Lawyers & Judges.

Sommers, M.S. (2006). Evaluating voice-based measures for detecting deception. *Journal of Credibility Assessment and Witness Psychology, 7*(2), 99-107.

STIG. (2008). *VSA-2000 lie detection lab*. Available: http://www.secintel.com/p-509-vsa-2000-lie-detection-lab.aspx [accessed February 2010].