Planning Support for Running Large Scale Exercises as Learning Laboratories

DISSERTATION

Presented in Partial Fulfillment of the Requirements for the Degree Doctor of Philosophy in the Graduate School of The Ohio State University

By

Martin Gregory Voshell, B.A., M.S.

* * * * *

The Ohio State University

2009

Dissertation Committee:  Approved by

David D. Woods, Adviser  Adviser
Philip J. Smith
Emily S. Patterson

Industrial and Systems Engineering
ABSTRACT

In many mission critical work domains effective scenario based training and observation are crucial to the success of complex socio-technical systems. Organizations employ many different approaches toward conducting these kinds of training sessions, but as recent high surprise disasters such as the Columbia loss and the 9.11 terrorist attacks indicate, there will always be new surprises that put that learning efficacy to the test. Anomalies will occur, new failure conditions will challenge existing systems and organizations, and these challenges of adapting to surprise and to resilience are nowhere more evident than in these mission critical domains. However, a significant amount of the training conducted in these types of organizations is not about training to be surprised, rather, its about showing individual competency and that current training is effective. These large-scale socio-technical organizations could be more resilient if they effectively exploit the opportunities from these exercises to capture learning and facilitate a deeper understanding behind the cognitive work in the domain.

This thesis proposes the *the learning laboratory* as a support framework for such exercise design. The learning lab framework serves as a general abstraction of CSE staged world study design and envisioning techniques and extends these approaches to cope with new scalability challenges to resilience. CSE has a long history of conducting research in complex domains utilizing effective staged and scaled world design techniques to support and illustrate the critical cognitive challenges of practitioners
at work. The learning lab incorporates a variety of these techniques into a common framework that can be applied to a variety of different types of exercises already being conducted in order to maximize organizational understanding and learning from scenario based observation exercises.

The initial learning laboratory framework synthesizes historical Cognitive Systems Engineering staged and scaled world design practices with findings from personal observations in three exercises: a tactical military intelligence scenario, an urban fire-fighting terrorist bombing, and a large-scale flu pandemic and terrorist attack. This synthesis exposed a number of challenges to successful learning from exercise design. These challenges were explored further in a series of critical incident interviews with a variety of researchers and practitioners in multiple domains. The insight gained from these interviews highlighted a number of decision-difficult design challenges as well as a number of support strategies for coping with specific challenges that must be planned for to conduct effective staged/scaled world studies as learning laboratories. These challenges were translated into support requirements for the design of an exercise planning tool prototype guided by the learning lab framework. Through a better understanding of the vulnerabilities that can stall learning opportunities and make conducting an exercise difficult, researchers, stakeholders, and practitioners responsible for future exercise design should consider the potential impact of taking such exercise design approaches to maximize learning yield and develop a deeper understanding of cognition at work.
For Pam, whose love and support has made this possible . . .
ACKNOWLEDGMENTS

I CAN'T IMAGINE A MORE STIRRING SYMBOL OF MAN'S HUMANITY TO MAN THAN A FIRE ENGINE.
VITA

March 3, 1980 .......................... Born - Thousand Oaks, CA USA

2002 ................................. B.A., Psychology, Skidmore College - Saratoga Springs, NY

2005 ................................. M.S., Cognitive Systems Engineering, The Ohio State University Columbus, OH

2005-2006 .......................... Visiting Researcher, Man-Machine Interaction Group, Technische Universiteit Delft, the Netherlands

2006-2008 .......................... Research Engineer, Department of Industrial and Systems Engineering The Ohio State University Columbus, OH

2008-present ...................... Senior Cognitive Engineer, Resilient Cognitive Solutions Pittsburgh, PA

PUBLICATIONS

Research Publications


**FIELDS OF STUDY**

Major Field: Industrial and Systems Engineering  
Area of Study: Cognitive Systems Engineering  
Minor Field: Information Analysis  
Minor Field: Cognitive Science
# TABLE OF CONTENTS

<table>
<thead>
<tr>
<th>Abstract</th>
<th>ii</th>
</tr>
</thead>
<tbody>
<tr>
<td>Dedication</td>
<td>iv</td>
</tr>
<tr>
<td>Acknowledgments</td>
<td>v</td>
</tr>
<tr>
<td>Vita</td>
<td>vi</td>
</tr>
<tr>
<td>List of Tables</td>
<td>xii</td>
</tr>
<tr>
<td>List of Figures</td>
<td>xiii</td>
</tr>
</tbody>
</table>

## Chapters:

1. **Introduction** | 1 |
   - 1.1 Research Goal and Organization | 6 |
   - 1.2 A Domain Description: Exercises in Crisis Response | 8 |
     - 1.2.1 The Building Block Approach to Exercise Design | 12 |
     - 1.2.2 Exercise Category Dimensions | 13 |
   - 1.3 Cognitive Training and Evaluating Human Performance | 21 |
     - 1.3.1 Staged World Design: Envisioning and Scaling | 25 |

2. **Three Learning Laboratory Observations** | 37 |
   - 2.1 Overview and Rationale | 37 |
   - 2.2 Stance, Limitations, and Scope | 39 |
   - 2.3 Ft. Huachuca | 43 |
     - 2.3.1 Overview and Scaling Dimensions | 43 |
     - 2.3.2 Exercise Details | 45 |
     - 2.3.3 Designing For Coordinated Observational Agility | 49 |
<table>
<thead>
<tr>
<th>Section</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.3.4 Lessons Learned</td>
</tr>
<tr>
<td>2.4 Kings Mall- Urban Firefighting Exercise</td>
</tr>
<tr>
<td>2.4.1 Overview and Scaling Dimensions</td>
</tr>
<tr>
<td>2.4.2 Kings Mall From Tabletop to Boots on the Ground</td>
</tr>
<tr>
<td>2.4.3 Observation and Analysis</td>
</tr>
<tr>
<td>2.4.4 Lessons Learned- Scaling at the Right Levels</td>
</tr>
<tr>
<td>2.4.5 Areas to Sustain and Improve</td>
</tr>
<tr>
<td>2.5 Recap of Observed Learning Lab Themes</td>
</tr>
<tr>
<td>2.6 Strong Angel III</td>
</tr>
<tr>
<td>2.6.1 Overview- A Really Big Show</td>
</tr>
<tr>
<td>2.6.2 Scenario- A Touch of Real-World Complexity</td>
</tr>
<tr>
<td>2.6.3 Scenario Execution- How it broke down</td>
</tr>
<tr>
<td>2.6.4 Lessons Learned</td>
</tr>
<tr>
<td>2.7 Synthesis Across Field Observations</td>
</tr>
<tr>
<td>2.7.1 Mapping From Test to Target</td>
</tr>
<tr>
<td>2.8 An Initial Framework for Learning Lab Design</td>
</tr>
<tr>
<td>2.8.1 Key Roles: Stakeholders, Observers, and Practitioners</td>
</tr>
<tr>
<td>2.8.2 Staging, Set-Up, and Functions</td>
</tr>
<tr>
<td>2.8.3 The Run-Time</td>
</tr>
<tr>
<td>2.8.4 Analysis, Learning, and Understanding</td>
</tr>
<tr>
<td>3. A Collection of Critical Incidents</td>
</tr>
<tr>
<td>3.1 Critical Incident Interviews</td>
</tr>
<tr>
<td>3.1.1 Participants</td>
</tr>
<tr>
<td>3.1.2 Materials</td>
</tr>
<tr>
<td>3.1.3 Procedure- Collecting the Data</td>
</tr>
<tr>
<td>3.2 Results</td>
</tr>
<tr>
<td>3.2.1 Case 1: The New Designer</td>
</tr>
<tr>
<td>3.2.2 Case 2: Designing the Right Sized Hole</td>
</tr>
<tr>
<td>3.2.3 Case 3: My Three Types of Exercises</td>
</tr>
<tr>
<td>3.2.4 Case 4: The Dayton EOC Tornado Exercise</td>
</tr>
<tr>
<td>3.2.5 Case 5: Ariane 501 Intelligence Analysis CTA</td>
</tr>
<tr>
<td>3.2.6 Case 6: The Failed NASA Aviation Study</td>
</tr>
<tr>
<td>3.2.7 Case 7: US Army Approach to Division Training</td>
</tr>
<tr>
<td>3.2.8 Case 8: Observation Support in the OR</td>
</tr>
<tr>
<td>3.3 A Need for Planning Support</td>
</tr>
<tr>
<td>4. A Learning Laboratory Planning Support Tool Design Seed</td>
</tr>
<tr>
<td>4.0.1 Scope</td>
</tr>
<tr>
<td>4.0.2 Analysis and Design</td>
</tr>
</tbody>
</table>
### LIST OF TABLES

<table>
<thead>
<tr>
<th>Table</th>
<th>Description</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>2.1</td>
<td>The five basic dimensions underlying a conceptual learning laboratory.</td>
<td>38</td>
</tr>
<tr>
<td>2.2</td>
<td>Huachuca as learning lab element overview. Instances of observation are marked underneath the five learning lab concepts.</td>
<td>59</td>
</tr>
<tr>
<td>2.3</td>
<td>Kings Mall as learning lab element overview. Instances of observation are marked underneath the five learning lab concepts.</td>
<td>76</td>
</tr>
<tr>
<td>2.4</td>
<td>Strong Angel 3 as learning lab element overview. Instances of observation are marked underneath the five learning lab concepts.</td>
<td>96</td>
</tr>
</tbody>
</table>
# LIST OF FIGURES

<table>
<thead>
<tr>
<th>Figure</th>
<th>Page</th>
</tr>
</thead>
<tbody>
<tr>
<td>1.1     Organizational learning is critical for creating foresight to cope with the complexity of unexpected crisis events in the 21st Century threat environment.</td>
<td>2</td>
</tr>
<tr>
<td>1.2     The DHS/FEMA Building Block approach to exercise and training design.</td>
<td>14</td>
</tr>
<tr>
<td>1.3     The key exercise planning roles and tasks in a typical emergency response exercise design.</td>
<td>17</td>
</tr>
<tr>
<td>1.4     The training committee assembles multiple functions together into a MSEL which guides simulator design, evaluator criteria, and controller planning in an exercise.</td>
<td>19</td>
</tr>
<tr>
<td>2.1     Five CSE Research observers were able to simultaneously watch the same events while communicating their observations with each other, and with a command center. Instructors watched and interacted with each analyst team individually, while researchers observed the interactions across all instructor groups and within each team. For each analyst team, the scenario and challenges faced played out slightly differently at run-time.</td>
<td>50</td>
</tr>
<tr>
<td>2.2     Example display of an observation command center observer receiving distributed team notes and a live video feed of a command briefing. This served as the raw data for analysis.</td>
<td>53</td>
</tr>
<tr>
<td>2.3     Process trace maps created from the raw data for our initial analysis.</td>
<td>54</td>
</tr>
<tr>
<td>2.4     The annotated timeline for one of the analyst teams depicts the relationship between observers, stakeholders, and practitioners against the events of the Northern Star scenario.</td>
<td>56</td>
</tr>
</tbody>
</table>
2.5 Annotated exercise staging overview relaying how planning goals led to events in the tabletop scenario, which were then complemented by additional events during exercise run-time.  
2.6 Detailed Process Trace analyses from the Kings Mall observation.  
2.7 Annotated animated timeline from the process created from the process trace data.  
2.8 Mapping the fire department’s organizational structure on top of a functional abstraction to identify potential support areas. The full functional analysis, right, was developed based on the process trace and led to cognitive descriptions of work.  
2.9 SA-III: Timeline of planned events, and run-time of actual exercise.  
2.10 A longshot of the initial learning laboratory framework consisting of three phases: staging, execution and run-time, and learning.  
2.11 Stakeholders, observers, and participants form the general stakeholder groups.  
2.12 In the staging phase, stakeholders define key purposes and functions for an exercise and then construct an exercise scenario incorporating multiple learning lab functions for the run-time.  
2.13 During run-time, participants experience the scenario events along with the learning lab functions as organic elements built into the normal flow of activity.  
2.14 At the end of an exercise the framework serves as a map of events from which one can process trace and evaluate learning shifts from local, to sharable, to generalizable.  
3.1 Overview of the frequency count tables and their scatter plot representation for capturing learning lab data.  
3.2 Flanagan’s original critical incident method.  
3.3 Categories observed in the New Designer case.
3.4 Categories observed in the ‘Right Sized Hole’ case. 137

3.5 Categories observed in the ‘My Three Types of Exercises’ case. 141

3.6 Categories observed in the ‘The Dayton EOC Tornado Exercise’ case. 146

3.7 Categories observed in the ‘Ariane 5 Intelligence Analysis CTA’ case. 150

3.8 Categories observed in the ‘Failed NASA Aviation Study’ case. 154

3.9 Great Louisiana Maneuvers Marker. Photo Copyright flickr user: Courtshouselover 156

3.10 Categories observed in the ‘US Army Approach to Division Training’ case. 160

3.11 Categories observed in the ‘Observation Support in the OR’ case. 164

4.1 The learning laboratory progression from outline to output. 171

4.2 The Learning Lab Designer being launched in an envisioned workspace. 174

4.3 Design Launcher Window: from here the user can choose whether to plan a new exercise, view an existing exercise, or explore the tutorial assistance. 175

4.4 The Learning Laboratory Video Tutorial Library allows the user to explore the CSE concepts and application interactions. 176

4.5 The Learning Laboratory Tutorial playing a video clip of Dr. Woods explaining key envisioning concepts for the planning of a learning lab. 177

4.6 All available exercises are represented by a high resolution icon and brief title. 178

4.7 The user has chosen to view Emergency Response learning labs and has selected the Kings Mall observation. 179
4.8 The Playback Viewer enables the user to explore the learning laboratory from a variety of perspectives to examine different stages of the exercise, how relationships played out, and how run-time challenges mapped to initial goals and learning.

4.9 Artifacts such as this Human Factors paper can be linked to the learning lab document and then be opened on the fly.

4.10 Starting a new plan in the main Learning Lab Designer window.

4.11 The user places the various stakeholder groups onto the learning lab ‘stage’ and fills out their initial goals prior to scenario generation.

4.12 As stakeholders work with one another to incorporate multiple goals and purposes, the actors for the event can be identified and included in the staging.

4.13 Stakeholders start the staging of the scenario by creating the exercise narrative. After the narrative is established, the sequence of events and learning lab functions are developed in parallel.

4.14 As the sequence of events are laid out, they can be visually linked to the learning lab functions as well as initial goals and purposes to provide a hypothesized roadmap of the scenario.

5.1 Looking at a cross-section of the learning laboratory upon exercise completion, one can now visually trace and see alignment between stakeholder objectives and actual outcome.
CHAPTER 1

INTRODUCTION

In many mission critical work domains effective scenario based training and observation is crucial to the success of complex socio-technical systems. Organizations employ many different approaches toward conducting these kinds of training sessions, but as recent high-surprise disasters such as the Columbia loss and the 9.11 terrorist attacks indicate, even high reliability organizations can find themselves in a response where training, tactics, procedures do not exist. Anomalies will occur, new failure conditions will challenge existing systems and organizations, and these challenges of adapting to surprise and to resilience are nowhere more evident than in these mission critical domains. However, a significant amount of the training conducted in these types of organizations is not about training to be be surprised, rather its about showing individual competency and that current training is effective. These large-scale socio-technical organizations could be more resilient if they could use the opportunities of these exercises to more effectively capture learning to facilitate a deeper understanding behind the cognitive work in the domain. This thesis presents a new Cognitive Systems Engineering (CSE) based approach that can be applied to the types of exercises already being conducted to maximize organizational understanding.
and learning from scenario based observation exercises in the form of the *learning laboratory approach*. 

This problem, how to become a more resilient organization by learning from exercises, is one held by by organizations in many large scale socio-technical systems such as military combat, planning and command and control, disaster response, nuclear control room operations, strategic intelligence analysis, space mission control, emergency department coordination, air traffic logistical planning, and network analysis and infrastructure monitoring, and many others. In general, it is often difficult to adequately explain such abstract CSE principles without showing how they instantiate in the context of rich, complex, real-world domains. Because of this, the patterns behind the learning lab approach are explored through lessons learned from multiple detailed observations in emergency management and disaster response operations. However, these general CSE patterns and the learning lab approach can be applied, with some differences in the details, to the host of domains mentioned above.

Figure 1.1: Organizational learning is critical for creating foresight to cope with the complexity of unexpected crisis events in the 21st Century threat environment.

Emergency management and crisis response serve as rich natural laboratories for the study of coordination and resilience in joint cognitive systems (JCS) and are
ideal test cases for exploring this new approach. Crisis response holds a unique place amongst the safety critical domains seen in CSE and Human Factors Engineering (HFE). One of the basic tenets of CSE is that effective organizations are prepared to be surprised (Hollnagel & Woods 2005, Woods 2005), and one of the primary mechanisms for how this could be achieved is through large scale exercises. The scale and frequency of recent natural disasters and terrorist events illustrate the growing complexity faced by practitioners and first responders operating today. Natural disasters such the 2004 South Asian Tsunami and Hurricane Katrina, ongoing conflicts in Africa and the Middle East, the London Underground bombings and the 9.11 terrorist attacks in New York City illustrate the diversity of these recent global crises. The last two examples in particular emphasize the changing threat environment faced by responders and response organizations (Lambakis, Kiras, & Kolet 2002, Hintze 2008, Voshell, Trent, Prue, & Fern 2008). The realities of the 21st century threat environment emphasize that the response community is facing new types of demands which challenge organizations’ abilities to be able to prepare, respond, and recover from crisis events and disasters. For effective and resilient future operations, enhanced planning and coordination among the various civil and military groups that could be involved in a crisis response is crucial. As a result of these dynamics, organizations have increasingly looked to seize upon the opportunities afforded by exercise simulation and new technology to simulate, explore, and test new ways to support and envision response processes for future emergency operations (Hintze 2008, R. Smith 2003, Baldwin 1994, Hintze 2008). Unfortunately, coupling technology with support in these types of complex domains poses significant challenges, for example, as described by noted rescue robotics pioneer Robin Murphy:
“Our experience is that every response is totally different and causes unforeseen problems or opportunities. We have never gone to an actual response and used the equipment the way we thought we would.”

(Murphy & Burke, 2005)

To successfully envision what might be useful for future operations, exercise planners and researchers must develop an understanding of the fundamental cognitive work in the domain. Exercises serve as simulations of envisioned futures to explore possible responses to events. As such, exercises must capture the cognitive complexities that crises pose for responders. As they are run today, exercises serve multiple purposes for multiple stakeholders. These exercise purposes and goals include training responders, providing drills and practice opportunities, affording evaluation of new systems, building trust across diverse groups, identifying critical decisions, enhancing coordination, and testing plans in light of new threats. Many organizations such as TEEX’s Disaster City, Picatinny’s Testbed EOC, HELP’s Calamityville, as well as large scale volunteer efforts such as the Strong Angel and Golden Phoenix demonstrations have been increasingly trying to use exercises to expand and share organizational learning. These latter types of exercises’ focus on learning resembles a more research oriented efforts that complements training looking at how future disasters could challenge response organizations, explore how new technology could support better team work across multiple groups, and test novel response practices that could transform response capabilities. Such efforts are indicative of explorations by the response community as a whole searching for new ways to be prepared. This sort of organizational learning is critical for creating the necessary foresight to cope with the complexity of unexpected events (Woods, 2004). This ability to effectively learn and adapt serves as a hallmark and significant challenge to any resilient or High Reliability Organization. HROs must
be able to anticipate the vulnerabilities to failure before capacity is stretched beyond return (Weick, Sutcliffe, & Obstfeld 1999). Joseph Pfeifer, Fire Department of New York (FDNY) Chief of Counterterrorism and Disaster Preparedness, stresses that “leaders must move beyond traditional reactive behavior by demanding resilient and adaptive approaches for managing complex incidents at both strategic and operational levels” (Pfeifer 2005).

In the modern crisis management environment, these exercises are no longer being conducted and evaluated for just training purposes. Instead, they are being run as complex simulations designed to fulfill a multitude of learning goals across a spectrum of cognitive work from skill based drilling to complex decision making under uncertainty. Designing exercises that meet these varied learning purposes has proven to be quite difficult given the design challenges posed by the changing environments, increased scales of events, logistical burdens, and uncertainty with training methodologies. Over recent years, CSEL has observed these difficulties plague crisis response exercises first-hand as participants and observers in large scale exercises ourselves (e.g., in search and rescue robotics training, in Strong Angel III, and in urban fire fighting operations). While a number of these difficulties represent classic challenges faced whenever researchers utilize training simulations with actual operators in experimental studies, there are a number of leverage points CSE can support for more effective learning yield and design. CSE has faced a number of these challenges in its own history conducting research and designing staged world exercises to support and illustrate the cognitive challenges of practitioners at work in complex domains. CSE has a rich history abstracting, supporting, and sharing knowledge from anomaly response in complex domains such as nuclear power, NASA mission control, and aviation
automation. Given the increased scale and the large number of stakeholders with various learning goals in modern large-scale crisis management exercises, a number of unique challenges emerge compared to this previous CSE work. The overlap between CSE and the natural laboratories of crisis management represents a profound support opportunity to guide decision-centered approaches to effective exercise design in crisis management.

To support discovery and effective learning in large scale exercises, this thesis proposes the learning laboratory approach as a CSE design framework to guide the creation of and steer effective learning output from runnable exercises. By planning for learning opportunities and by anticipating scaling challenges, the learning lab design approach enables multiple stakeholders to plan, collect data, and evaluate performance trends to support a variety of abstract goals and learning purposes. Conceptually, a learning laboratory is an exercise based on a collection of CSE scaled/envisioned world principles that results in local, sharable, and generalizable learning which contributes to the understanding of cognitive work in a domain. Expanding on that concept, the goal of this thesis is to explain:

- What CSE concepts define the learning laboratory approach.
- Why field exercises should be designed as learning laboratories.
- How to design field exercises as effective learning laboratories.

### 1.1 Research Goal and Organization

The central focus of this work is to define the critical support functions that must be planned for and built into runnable exercises to effectively support and explicitly
link learning purposes with analysis to lead to discovery in any complex domain. The learning lab concept extends CSE staged world methods to help organizers and researchers observe, identify patterns of work, and envision new leverage points and possibilities from exercises. This thesis synthesizes findings from field observations along with a series of interviews with emergency/anomaly response subject matter experts to critically assess how to conduct large scale exercises as learning laboratories. By taking a CSE approach to exercise design grounded in the principles of staged world envisioning, these findings will be used to create a design seed of a working database planning tool that supports the design and evaluation of human performance in future exercise designs. The learning laboratory approach aims to enhance the practical relevance, scientific credibility, operational effectiveness, and learning yield from current and future training design.

The remainder of this chapter reviews gaps in current approaches to disaster exercise planning and identifies how complementary CSE techniques can be incorporated to bridge these gaps. This complementary system grounded in CSE staged and envisioned world design forms the conceptual basis of the learning lab approach. The learning lab concept is further refined in Chapter 2 based on results from a meta-analysis of three field exercises evaluated as learning laboratories of different scales. These exercises include a found case working with a military intelligence training facility, a future incident case with a metropolitan fire department, and observation in a large-scale future technology international disaster response training consortium. Findings from the meta-analysis are synthesized to develop an initial Learning Laboratory design framework. Chapter 3 uses the design framework as guide for a series of critical incident interviews and retrospective walkthroughs with prominent CSE researchers.
and crisis response practitioners. Findings from the interviews reveal a number of challenges to planning and learning that exercise designers must cope with and overcome for successful learning to occur. These create new support demands and further characterize relationships within the learning lab framework. Chapter 4 translates these exercise design challenges into support requirements for a proposed exercise planning support tool design seed. Last, Chapter 5 discusses the learning laboratory as a general abstraction of staged world design and its potential to bridge the gap between learning laboratory research and the design of future systems.

1.2 A Domain Description: Exercises in Crisis Response

In some organizational cultures, training emphasizes adaptive and creative problem solving. The following disaster account serves as a success story where two mission critical domains, two organizations heavily invested in training cultures, were able to coordinate together in response to an escalating and complex disaster.

“So enough redundancy was built into the [hydraulic] system to where the odds were placed at 1 to 10 to the 9th power, or a billion to 1, that complete hydraulic failure would occur... Well, on July 19th, Murphy’s Law caught up with us, and we did lose all three systems... We had no ailerons to bank the airplane, we had no rudder to turn it, no elevators to control the pitch, we had no leading-edge flaps or slats to slow the airplane down, or help us slow down, once we were on the ground.”


In the mid-afternoon on July 19, 1989, United Airlines Flight 232 crash landed in Sioux City after an uncontained engine failure destroyed the DC10’s hydraulics system. As recounted in the above quote by pilot Capt. Al Halynes, this failure was statistically improbable and aircraft engineers were certain there was enough redundancy in the hydraulics system to prevent such an event from occurring. The
engine explosion sent shrapnel from the fan blade flying, severing all three of the independent hydraulic lines. Based on the flight crew’s successful coordination within the cockpit and with air traffic control, they were amazingly able to pull off a crash landing. Just as important to this success were the efforts of the emergency response crews on scene. With only 24 minutes warning, well practiced and recently trained rescue units and response agencies consisting of over 40 local fire agencies and 300 National Guard personnel were on the ground and ready. Their efficient response ultimately saved 186 lives in one of the few major crash landings of a large passenger jet in aviation history. Many of the first responders indicated that their preparedness was largely testament to the utility of a recent serendipitous large-scale training exercise.

Two years before the fateful UA 232 crash, Sioux City’s Disaster Services Center (DSC) conducted a full-scale exercise based on a commercial plane crash (FEMA IS 239, 2008). The DSC consisted of 40 stakeholder representatives from the various rescue and emergency response agencies in the community that was responsible for reviewing response procedures and planning ‘realistic’ exercises. The after action report from the full scale exercise highlighted a significant amount of confusion in communications as well as inadequate rescue apparatus and resources at the scene of response. Rigorous planning and re-development of response plans were initiated after the exercise to address the observed shortcomings. The practice and cooperation paid off. As quoted by the Assistant Fire Chief the day following the crash [DHS/FEMA 2009]:

“We made mistakes [in the simulation]. The mistakes we made then did not materialize yesterday.”
The evolving situation and cascade of events both in the air and on the ground is a remarkable story of adaptation to surprise. However, focusing on the design of the fortuitous disaster response exercise highlights some key differences between the training approaches behind each organization. First, a quick history lesson is needed to explain what led to these training differences. In typical aviation training 30 years ago, a majority of pilot training was focused on how to build the airplane (Billings, 1996). Training approaches and certification emphasized intricate technical knowledge of flight systems while assuming human factors, such as group dynamics, would come for free. This came to a head as an increasing number of high profile accidents occurred in the 1970’s while at the same time the industry faced increasing pressures to reduce training time- a classic “faster, better, cheaper” resilience tradeoff challenge (Woods, 2006). The aviation environment was impacted further with the rapidly escalating complexity introduced by new cockpit automation systems. Key decision makers could see the shifting environmental demands and began to realize these traditional training approaches were no longer sufficient. While aviation simulation technology continued to define the cutting edge of high fidelity training, the functional/operational fidelity (Caird, 1996) built into the training was severely absent. Approaches had to be redesigned to support pilots learning how to operate these complex systems rather than simply training them on how the machines operate so as to better cope when systems reached their limits. A new approach was advocated by key individuals that saw this mounting complexity and Cockpit Resource Management (CRM) was born. Combining classroom instruction with full-mission flight simulation CRM was developed as a new training approach that complemented traditional training programs but emphasized teaching teamwork and cockpit coordination techniques and not only
stick and rudder controls. Training that encouraged adaptive responses as a team likely contributed to the ability of the pilots in Sioux City being able to land. When starting to fly, pilots are inundated with the first principles of flight cementing why planes don’t fall out of the air- they are not simply trained by rote mechanical response and memorization. Because a good pilot understands the physics and aerodynamics of flight, complemented by higher level coordination and cognitive training, pilots learn how to functionally understand how to respond to an envelope of human-human and human-machine coordination challenges. Emergency response, however, is now in a similar situation as aviation 30 years ago facing a significant reckoning with complexity as the bellwether events from the beginning of this chapter indicate.

Sioux City worked thanks largely to proactive response leadership that had recognized a training need due to environmental conditions (proximity to an airport) and a lack of interagency coordination. Emergency response organizations held an exercise, drilled the scenario with multiple agencies, and adjusted what didn’t work. These types of exercises are largely doctrinal in nature. Doctrinal training instills the tactics, techniques, and procedures to fulfill particular tasks and is created based on the envisioned situations expected to be encountered. While such core skill-based competencies are critical (Pfeifer, 2005), to some degree emergency response training is still ‘learning up’ responders the same way aircraft pilots were trained as mechanics. As will be discussed below, these exercises are often designed to hone plans and procedures and drill responders into rote following. Reflective response practitioners (Schön, 1983) such as Chief Joseph Pfeifer, FDNY, recognize this as a problem and CSE researchers and practitioners have long warned about the potential brittleness when plans and procedures meet unanticipated variability in a dynamic and changing world (Woods &
Roth, 1988). These skills must be complemented by cognitive knowledge-based competencies, what Caird described as building the functional fidelity, which are necessary for improvisational and adaptive decision making when events move beyond prescribed doctrine. Training scenarios must be designed that are complex enough to challenge these competencies at both task and team levels as situations move from textbook into exceptional cases (Woods & Hollnagel, 2006; Woods, 1994). Illustrative of the escalation principle (Woods & Hollnagel, 2006), as problems cascade, uncertainties place more demand on the cognitive and coordinative work of the agents in the system to respond. This is a profound challenge, and in a personal interview with one of DHS’ exercise and planning managers, he echoed this growing concern in their own community regarding these types of challenges:

“The other challenge we have, we don’t really have well detailed plans. The problem associated with that, is one of, if you try to respond to every type of emergency, then your plan becomes very general. The more specific the emergency you’re responding to, have more specific plans, then what happens when it’s anomalous? We can’t deal with that.”

- C.D.R., personal interview

1.2.1 The Building Block Approach to Exercise Design

Response and rescue organizations train and conduct exercises for a variety of reasons to improve individual training and system capabilities. Sometimes training is part of regulatory requirements such as with the Nuclear Regulator Commission, or licensing to practice as with airline pilots and medical staff. The key difference though is how they train. The Federal Emergency Management Agency, FEMA, and the Department of Homeland Security, DHS, emphasize the driving force behind any exercise program is to use training exercises to improve an organization’s readiness and
capabilities. Exercise lessons give a means to evaluate plans, procedures and operations, cultivate teamwork, and illustrate an organization or community’s ability to prepare for disaster events. The main goals for conducting exercises serve to better establish roles and responsibilities, improve interagency coordination, identify resource gaps, hone individual performance skills, and identify improvement opportunities. Exercises may differ across a number of dimensions with regard to their scale, complexity, scope, purpose, approach, and cost. FEMA and DHS encourage that comprehensive exercise programs use a ‘building block approach’ (see Figure 2.1) where multiple exercises are planned in an iterative cycle. Each exercise is progressively more complex in the cycle and builds on the previous exercise as planners increase scale and complexity. This requires meticulous planning as each level of complexity integrates more sophisticated simulation, more preparation time, and more personnel. The building block approach consists of two general categories of exercise activities, discussion-based and operations-based exercises.

1.2.2 Exercise Category Dimensions

Discussion-based exercises are guided or facilitated/moderated discussions that do not involve the deployment of resources and as such are the least complex type of exercise to run. These exercises are generally used when problem holders wish to evaluate or focus on higher level issues such as strategy, policy, general training and procedures, or relationship building with other agencies. An example of a discussion-based exercise objective would be evaluating the standard operating procedure for a specific type of response. There are three types of discussion-based exercises- seminars and workshops, table tops, and role play gaming. Seminars and workshops are used
Figure 1.2: The DHS/FEMA Building Block approach to exercise and training design.

to familiarize participants with plans, procedures, and equipment. These are led by a facilitator and do not involve any simulation. A tabletop exercises is a facilitated analysis of emergency events in a low-stress environment. There is an element of simulation, but its power is based on participant discussion and problem solving over the course of events. A facilitator is necessary to lead discussion, inject information, and guide problem solving. Tabletops tend to lack realism and do not stress the response participants. The last type of discussion-based exercise is role-play gaming or tactical decision games. This form has been gaining in popularity in recent years especially as a means for scenario planning and to evaluate decision support systems (Turoff).
Role-playing exercises involve a game-master who creates conditions for play behind a scenario that participants assume roles in. The role playing is a collaborative exchange between the game master who creates and provides contextual details around scenario conditions and obstacles. This allows for more dynamic scenarios in simulation while still in a low stress environment.

Operations-based exercises do involve resource and personnel deployment and represent an order of complexity above discussion-based exercises. An example of a operations-based objective would be testing the ability of a specialized HAZMAT team to coordinate with a local response agency during a chemical incident. Evaluation would then look at such objectives as the effectiveness of communications, how well equipment was used, or how well different agencies coordinate and cooperate. Operations exercises come in the form of drills, functional exercises, and full-scale exercises. Drills are single-agency and usually run with a specific operation focus. They are commonly used to train on new equipment, test new procedures or maintain and practice skills. Drills provide immediate feedback but in a relatively isolated environment and require a supervisor or exercise manager with a support staff. The next level up in complexity are functional exercises, also known as command post exercises in military settings. These are aimed at either single or multi-agency activities that test capabilities and functions in a simulated response. These exercises are fully interactive simulations, are high stress, and require rapid decision making under time pressures. Often conducted in Emergency Operations Centers, functionals test multiple functions of an organization’s operational plan. These tend to be complex and primarily geared at testing coordination. Functionals also require a controller to manage and direct
the exercise, participant players to assume typical roles, confederates or computer agents for simulated roles, and evaluators to monitor and assess performance. Last, full-scale exercises try to closely simulate real events under high fidelity response conditions. These are aimed at multiple agencies and jurisdictions which involve the coordinated deployment of personnel and resources. These high stress realistic exercises follow a scripted exercise scenario and enable organizers to assess plans and procedures as well as evaluate a response under realistic crisis conditions. Full-scale exercises represent the most complex logistics and require a dedicated planning team and multiple controllers and evaluators that must coordinate the large number of personnel and equipment.

**Key Roles**

The design of an exercise is headed by an exercise design team leader with the support of an exercise planning team. This team is responsible for the development, conduct and evaluation. The team defines objectives, develops scenarios, writes documentation, and holds briefings and training sessions. The planning team consists of key participating agencies and organizations. Exercise design begins with need assessments to identify what capabilities should be exercised. The scope of the exercise defines the kind of exercise participants involved. The team must then establish the purpose of the exercise, often in the form of a need statement, as well as objectives based on the descriptive performance expected from participants. The purpose and objectives lay the groundwork for the scenario development as well as evaluation criteria.
Figure 1.3: The key exercise planning roles and tasks in a typical emergency response exercise design.

Basic Planning Functions

The scenario is defined as the story-line that drives the simulated emergency for the exercise. The scenario is based on the scope of the exercise and provides the context and details of the story which sets up conditions that allow for the evaluation of capabilities. Current guidelines encourage scenarios to be threat based and performance based, realistic, and challenging but not too challenging to overwhelm the participants (this ‘designed for success’ trend will be discussed later). Feedback is often elicited at the end of exercises based on feedback forms, hot-washes immediately following the exercise (hotwashes function and group groundings to develop consensus and identify commonalities and areas for improvement), and final debriefings between planners, facilitators, controllers, players, and evaluators. The first step of scenario design is composing a narrative. The narrative is a brief description of events leading to the
start of the exercise. This establishes the reason participants are involved and also sets
the stage for ensuing action. After the narrative, major and detailed events are fleshed
out and linked to the observable actions desired. This depends on detailed scripting
and clear objectives. The series of events are laid out along with their expected actions.
Last, messages are prepared that communicate event information across the various
stages of the simulation. These may either be pre-scripted messages or geared toward
specific response actions. Last, all the events, their times, and expected actions are
put together as a Master Scenario of Events List (MSEL) to help guide controllers,
evaluators, and simulators.

There are a variety of functional roles required for both discussion and operations
based exercises. In operations-based settings, controllers and evaluators keep exercises
on track and are responsible for evaluating performance. The controllers plan and
monitor play, operate the exercise site, and occasionally take on absentee organization
roles. Evaluators specifically track observations in relation to evaluation objectives
and assist in result analysis. Exercise evaluation is based on how well objectives were
achieved and how to improve future exercises. The evaluators, being responsible for
observation, gauge performance against expected outcomes and determine how to
make changes to guarantee the desired outcomes. The DHS and HSEEP evaluation
process consists of 8 steps and looks at three levels of evaluation: task level abilities,
organizational level, and mission level.

1. Plan and organize the evaluation

2. Observe the exercise and collect data

3. Analyze data
Figure 1.4: The training committee assembles multiple functions together into a MSEL which guides simulator design, evaluator criteria, and controller planning in an exercise.

4. Develop the draft After Action Report (AAR)

5. Conduct an After Action Conference

6. Identify improvements to be implemented

7. Finalize the AAR and Improvement Plan (IP)

8. Track implementation
The After Action Review (AAR) is one of the most critical artifacts for learning to come out of this process and serves to summarize the detailed analysis of the exercise. The AAR is assembled by the evaluation and planning teams based on evaluator observations, evaluation guides, hot washes/debriefs/feedback forms, and organizational plans and procedures. The goal of the AAR is to descriptively evaluate events of the exercise, highlight adaptations and issues that arose, and prescriptively recommend improvements. The AAR serves to provide feedback both to the participants on their performance, as well as feedback to the planners as to the effectiveness of the exercise itself. The AAR recommendations and lessons learned are then integrated into an improvement plan to future performance. In military team performance evaluations, a significant amount of effort is placed on supporting trained observer roles who each watch and critique an exercise participant and rate performance. It is not uncommon to have a one:one mapping of evaluator to exercise participant which becomes quite challenging as observations scale in size and the need to correlate events across observers increases (Campbell, Freeman, & Hildebrand, 2000). Another approach is using a single trained observer to evaluate an entire team and then provide performance scores after specific segments of activity, similar to how CRM in aviation training is conducted (Helmreich, Butler, Taggart, & Wilhelm, 1995). Again, scaling makes this very difficult to track- and depending on how observations are recorded, most often with notes, there’s no way to go back and re-analyze events. AAR debriefings are critical to instruction and learning, these replays trigger reflection. As outlined by Wiese, Freeman, Salter, Stelzer, and Jackson (2008), the AAR should function to make sure the right participants learned the right lessons so that they will recall correct and incorrect performance and the context around it so that learning transfer is either
encouraged or discouraged in future situations. A good AAR will help participants discover performance failures and successes, reflect on them, and diagnose causes. Last, they enable more social evaluation of performance and competence across other participants. What they do not do, however, is abstract to generalizable learning. It would be rare where a developer or researcher could take an AAR and immediately be able to interpret support requirements or immediately perceive cognitive work insights. While this process can be very effective for local and sharable learning, it is this further abstraction toward generalizable learning at the understanding level that from a CSE perspective would be the most critical to develop.

This overview serves as a general end to end breakdown of current exercise planning in emergency response. Many agencies use this prescriptive process (and often times Federal funding depends on its usage) to develop exercises for crisis response and emergency management training, evaluation, and testing. However, this thesis argues that there are a number of decision support and learning gaps in this approach that do not easily translate into building understanding so as to guide future system design and lead to new approaches. This necessitates CSE guidance to develop this understanding through cognitive problem solving and decision making skills to address multiple learning goals. This general outline for beginning exercise planning can be extended into a learning laboratory by scaffolding CSE guidance and support on top of it to extend its learning potential.

1.3 Cognitive Training and Evaluating Human Performance

“It is dangerous to conclude that actual training has occurred simply due to exposure to a scenario. In fact, team training requires a thoughtful specification of which knowledge, skills, and abilities need to be trained
and careful attention to and debrief mechanisms that foster learning and skill acquisition.” (Salas, Bowers, & Cannon-Bowers, 1995)

Gaining insight into how people actually perform work, as opposed how they are supposed to perform work, lends itself to multiple approaches toward supporting the design of CSE grounded response training programs. This starts by looking at the learning process itself. A significant amount of training in response domains is based on traditional classroom teaching. Classroom approaches focus on doctrinal practice and running through anticipated emergencies as well as repeated drilling for existing plans and procedures. However, a large degree of job-related learning develops in less traditional ways. Such ‘situated learning’ in the form of apprenticeships and co-worker teaming are quite often found in emergency response and similar domains. Firefighters, for example, make frequent life and death choices and must be well versed in critical decision making (Klein, Calderwood, & Clinton-Cirooco, 1986; Klein, 1993; Hintze, 2008). When working with the U.S. Marine Corps on cross-training decision making approaches, Vincent Dunn, Deputy Chief FDNY (Ret.) imparted how fire department chiefs effectively train officers (Dunn, 1999). Dunn highlights interesting similarities and differences between the two domains. In the military, the decision making often ends when the war ends. When a new conflict arises, those previous decision makers are often no longer in the same roles or are retired. In contrast, the standing fire service membership “fights a war that never ends...” (Dunn, 1999), gaining experience through study and practical application on a day to day basis. Fire leadership comes from within the organization, every chief was once an officer and every officer was once a firefighter. Experience in fire training, like in the military, is built upon mentorship systems. They refer to this as “growing their own”. Across all echelons in the fire
department, staff is coupled with veteran expertise to support the positive diffusion of learning. Similarly, the military has long developed and valued field exercises and simulation programs for unit training that aims to improve individuals’ abilities to work together as well as strengthen inter-organizational coordination (Alberts and Hayes, 2006). Doctrinal training encourages mastery of the tactics, techniques, and procedures to fulfill particular tasks and is created based on the envisioned situations expected to be encountered. As was first mentioned in the discussion of Sioux City, while rote proceduralization and core competencies are critical for emergency responders (Pfeifer, 2005), so are the cognitive skills (e.g., situation assessment, planning) necessary for decision making in situations where events move beyond prescribed doctrine where formal procedures may not exist. In studying crisis response in offshore oil and gas installations, Flin and Slaven (1996) found that using methods grounded in cognitive work, in this case rapid primed decision making and cockpit resource management, provided a much better grounding for a small-scale simulations to train decision making under emergency conditions as opposed to formal decision analysis approaches. Similar analysis by Mumaw, Swatzler, Roth and Thomas (1994) looked at cognitive skill and decision-making requirements in complex nuclear accidents which reinforce the validity and critical need for such approaches. More recently this has been explicitly looked at in regard to scaling the complexity of evaluation scenarios (Patterson, Roth, & Woods, in preparation). Both of these cases largely contrast typical current approaches to emergency response training. This leads to a significant semantic difference between training for emergencies versus training for crises. An emergency is defined as an abnormal situation where proper actions are based on defined tactics, techniques, and procedures (Barnett, Gatfield, & Habberley, 2002). In a crisis on the other hand,
successful actions may not be based on established procedure (and might also conflict with it) and will require to be thoughtful and adaptive in their response. This is analogous to a general pattern of work seen in CSE- anomaly response (Woods & Hollnagel 2006). Thus, while training for emergency conditions can be thought of as drilling in plans and procedures, as with the Sioux City exercise, training for crises requires additional training for higher level cognitive skills such as decision making, problem solving, and situation assessment. By seeing this as a kind of anomaly response, training then can be focused on established patterns of what makes anomaly response difficult (i.e. recognition, expectation, attention control, diagnosis, and re-planning (Woods & Hollnagel 2006)). In order to emphasize these skills as well as inform future design, a cognitive understanding of the domain must be built and then researchers can re-iterate and learn the new ways these challenges play out with practitioners at work.

Empirical investigation in CSE relies on a variety of task (CTA) and work analysis (CWA) methods in order to create a model understanding of a work domain from which one could base training. In order to build this understanding, researchers must gain insight and build up knowledge of the domain they are studying. This domain expertise enables them to identify and analyze the cognitive demands of the domain and see how practitioners respond to those demands. To do this, multiple CTA techniques and methodologies can be utilized. These can range from interviews, retrospective analyses of critical incidents, to field studies and field observations of natural, to simulated environments. The primary objective of cognitive training is to build operator and practitioners skills in coping with cognitively challenging events (Roth, Mumaw, & Pople 1993). Typical incidents involve multiple factors that make them unique and
cognitive training can better equip operators to handle these unique features (Woltjer et al., 2006). CSE has a long history of creating such exercises often being built as part of existing training programs. Designing such studies entails a conceptual shift away from traditional experimental design and statistical analysis methods and a movement toward staging and analyzing field observations and participant-centered work. To do this CSE borrows from a variety of research traditions which, at their heart, are grounded in shaping the conditions for observation in staged world design (Woods & Hollnagel, 2006).

1.3.1 Staged World Design: Envisioning and Scaling

Cognitive Systems Engineers study the world through a variety of techniques—all grounded in situated observation. In these observations, CSE practitioners can neither ignore the underlying work domains nor can they look only at static worlds. Cognitive systems are joint systems spread across human and machine agents in dynamic and changing worlds. The CSE researcher’s job is to discover meaning in these natural laboratories. Cognitive task and work analyses are successful to the degree that they uncover how agents’ behaviors (agents being used synonymously with humans, actors, or software agents) are adapted to constraints and demands in a field of practice (the environment). The field of practice is a relationship to the agents in it, the tools they use, and it is through these analyses and observations that new understanding about those relationships are revealed. It is this interplay between environment, agents, and tools that define the joint cognitive systems from which to observe and describe adaptive cognitive work, demands, and adaptations (Woods, Johannesen, Cook, & Sarter, 1994). CSE is typically concerned with those general
themes, these patterns of cognitive work and challenges to decision making that emerge from watching practitioners at work. This is what CSE looks for when creating and observing staged world studies in order to build understanding. Unfortunately, often when staged world studies are written up, methods sections tend to be sparse and they rarely focus on the difficulties and challenges encountered in the processes behind study execution. Behind all of these exercises, it is easy for researchers to lose track of such logistics and challenges as: balancing multiple stakeholders and practitioners with potentially conflicting and competing goals; difficulties and challenges to overcome going through the planning process behind an effective study run-time; challenges to successfully sharing the learning and gaining understanding from the studies across practitioners, stakeholders, as well as researchers. These components are valuable types of learning that must be preserved, captured, and reflected on as the multiple dimensions of planning support design are critically explored.

**Envisioned Worlds for Learning**

Designing new systems to advance training in crisis response requires envisioning future operations. CSE uses envisioning to predict and ultimately learn about innovation. This is not easy and is an ever-present challenge to cognitive task design and work analysis methods in general. As [Dekker and Woods (1999)](1999) explain, many CTAs are carried out based on the current field of practice- a small issue given that new technology changes the fundamental nature of the work. This prediction problem of envisioned worlds, referred to as the *envisioned world problem* ([Dekker & Woods, 1997](1997)), is based on four dimensions ([Roesler, Feil, & Woods, 2001](2001)). These dimensions have implications on planning design and therefore must be balanced to better anticipate
such work challenges. These dimensions include \textit{plurality, under-specification, groundedness, and calibration}. The first two, plurality and under-specification, can be seen when different groups and stakeholders all have different perspectives of what future operations will be like. If certain stakeholder or vendor groups are advocating certain technology, or if there is no technology representation, gaps and oversimplifications will emerge. Plurality specifies that there will be multiple views to describe the envisioned world and under-specification states that each of those views will be only a partial or oversimplified representation \cite{DekkerWoods1999}. There will always be competing and conflicting goals at the stakeholder level and technologist/practitioner gaps cannot always be filled and can create additional challenges at the actor/participant levels. On top of this, stakeholders responsible for exercise planning need to be able to balance the short-term immediate goals and functions of the exercise along with long-term goals of cultivating a learning environment and advancing research \cite{HoffmanWoods2000}- they must always keep the “big picture” in mind. The environment they are operating under must balance the local needs of the stakeholders and actors along with the general goals of adding to the research base.

The other two dimensions that must be addressed in envisioned world design are groundedness and calibration. Groundedness refers to how well predictions about the future are cemented in patterns of work from the research base. These operations need to be based on empirical research in order to help predict reverberations of change (i.e. patterns of work such as anomaly response from \cite{WoodsHollnagel2006} are a collection of candidate generic patterns). Lastly, calibration refers to the confidence, or overconfidence, that an envisioned system is feasible. It is critical that designers are open to revision and can acknowledge in the face of new observations or adaptation.
In the context of exercise and observation design, it is crucial to be able to create a general environment that allows system designers and stakeholders to collaboratively design for envisioned worlds. In any of these, there are multiple sets of stakeholders and multiple sets of participants that must be synchronized and coordinated such that they can be developed into a script for running a controllable exercise at the run-time. Stakeholder, researcher, and practitioner roles must all be identified from the start and planned for throughout the set-up, the run-time, and the sharing of learning. To do this requires a thorough understanding of the JCSs at work and this understanding must guide the creation of scenarios containing problems of interest and shape the conditions of observation as exercises are developed.

The envisioned world problem further highlights a critical difference between technology centered and practice-centered approaches to design that must be bridged. In order to create a collaborative environment to facilitate learning and coordination, this gap must be mitigated before going into the planning of an exercise design. System designers and exercise organizers must be able to contrast and bridge between both technologist and practitioner perspectives. Convergent CSE approaches emphasize that in order to design for the entire joint cognitive system, designers and planners must be able to understand the consequences of introducing new capabilities into the domain of practice.

The world is not static and neither is the environment being created for the exercise staging. To cope with this, CSE approaches advocate a number of converging methods including: P. J. Smith and Geddes, including Cognitive Task Analysis (CTA), Applied Cognitive Work Analysis (Elm, Potter, Gualtieri, Easter, & Roth [2003]), Cognitive Field Research and ethnographies [Hutchins [1995]], to capture the dynamic nature
of work and to better predict and meet new challenges. The methods employed
are intended to bridge the technologist-practitioner and other stakeholder gaps to
better understand how they benefit one another. This poses the initial challenge to
exercise and learning lab planning: the exercise(s) and scenario(s) being simulated are
envisioned systems themselves. Organizers and stakeholders consist of technologists
and practitioners responsible for planning the exercise and injecting points of change.
They need both a common understanding of the challenges to the domain of practice
going into the planning of an exercise, and they must be able to plan and adapt their
methods as they introduce or experience elements of change. This requires ongoing
analysis and observation of the organizers and the entire system as a whole as well
support for observation and analysis of the actors in-scene. This critical demand for
supporting observation will become a central theme played out throughout the rest of
this thesis.

Staged/Scaled World Design

Building scenarios and conducting studies for envisioned worlds is noteworthy be-
cause of the unique power designers have to control the world they have created. These
studies can range from empirical spartan labs and experimenter controlled simulations,
to direct observation in the field of practice, to staged/designed problems (Woods &
Hollnagel [2006]). Authenticity is the key to creating problems that both represent
the vulnerabilities and challenges that exist in practice coupled with how participant
problem-solving expertise is elicited in the study (Woods, Christoffersen, & Tinapple
[2000]). Studies in envisioned worlds span both of these authenticity dimensions by
creating hypothesized cognitive challenges to be faced by real practitioners. The
envisioned world puts stakeholders into a prediction role, requiring them to generate
hypotheses about future operations and new challenges, failure modes, and potential breakdowns faced in work domains. These then guide stakeholder purposes and goals as planning translates into exercise layout.

CSE tends to operate between the spectrum of staged/controlled studies and flat out observations in the wild. [Woods and Hollnagel] (2006) and Woods (2003) identify three families of methods that can be employed when studying Joint Cognitive Systems at work: natural history methods, experiments in the field, and spartan laboratory experiments. Natural history techniques alone, while they can provide valuable insight, cannot lead to generalizable models and patterns. More rigorous investigation requires a degree of empirical testing and this can be accomplished through staged world simulation. Staged and scaled world studies are experiments-in-the field that stage situations of interest through simulation in an experimental setting ([Woods & Hollnagel], 2006). Staged world techniques lie between anthropological natural history field observations and classic experimenter designed laboratory studies. In field observations, researchers have little control over the conditions they are observing. In traditional laboratory experiments, the ecological world is often over-simplified and critical aspects of the environment are often foregone for control over specific isolated factors. Like much of science, the staged world study falls between these two ends- a balance between data and theory, opportunity and control.

A variety of methods are available to approach discovery and learning across the three families of observation. These include building and analyzing corpuses, direct observation of work domains, designing cognitively challenging scenarios for simulation, staged and scaled world simulations, process tracing and protocol analysis techniques, collaborative envisioning through mock-ups and prototypes, cognitive
simulation, and functional analysis and goal decompositions of domain work. These approaches are often complementary and best when bootstrapped together to describe the most complete picture of work as possible incorporating multiple perspectives (Potter, Roth, Woods, & Elm, 2000). The power of staged world studies comes from stimulus sampling. Because the studies deal with real-world practitioners and designer hypothesized views on the challenges to practitioner work, researchers and designers intimately care about what problems practitioners receive. In classic experimental terms, there is the creation of a target situation and test situation and designers are interested in the mapping between them (Woods, 2003). It is from here one gets generalizability, the mapping from test to target, which designers can re-use and analyze if the same patterns are elicited. This explicit mapping is critical to be able to trace out. This traceability is a defining characteristic and becomes evident in process traces when it turns to evaluation. When designers create a scaled world study they control the context of the case and end up with the ability to observe and stage repeated observations. While classic experimental concepts such as random sampling or control groups are often not feasible in staged world scenario simulation, designers can create prototypical scenarios that allow them to compare and contrast performance across interrelated sets of scenarios utilizing the same cognitive pressure points (Woods & Dekker, 2000). Similarity across responses to events is often just as valuable as deviations and adaptations.

A number of CTA methods can then be used to conduct the staged world studies. The future incident technique, for example, is one particular CTA strategy to guide learning objectives and staged world study design. As defined by Dekker and Woods (1997):
“The future incident method is based on developing a failure or near miss that could happen given one view of how the envisioned world might work. The future incident is based on our knowledge of classic design errors – cases where new technology shaped cognition and collaboration in ways that produced problems such as clumsy automation, coordination surprises or other difficulties that contributed to incidents or accidents (e.g., Woods et al., 1994, chapter 5). The scenario designers look at different views of the envisioned world looking for places where design errors would negatively impact on cognition or collaboration of future practitioners. The resulting future incident is packaged in the format of actual incident reporting formalisms used in that field of practice.”

The future incident technique can be used as a story framework that forms the context around which to stage micro-scenarios and inject probe events. This could be used as an effective part of planning to bootstrap scenarios together (Potter et al., 2005). By working with practitioners using limited role-playing or walkthroughs, multiple perspectives on problems and cognitive challenges in the world can be considered and further developed. As such, the scenario design itself is treated as a collaborative envisioning process up front (Schoenwald, et al., 2005). Probe designs are created and refined to illustrate the hypothesized demands for a scenario designed around the complexities of the domain (Patterson, 2004). These scenarios then help designers and practitioners share and examine how to seize upon aspects and organize the exercise scenario around the actual flow of domain knowledge around challenges and disruptions. The scenario design process itself serves as an envisioning tool that both describes and can be used to share the complex interactions (and potential unanticipated side effects), stories, and adaptations in work from which to initiate feedback from practitioners, designers, and technologists.

Staged/scaled world studies can incorporate a range of scenario classes from full-scale exercises in training grounds, to functional exercises and computer simulation,
to thought and table-top exercises. These all require researchers to be able to design effective problems in order to understand and innovate systems that support situated cognition and collaboration in complex domains. By combining domain research and working with practitioners, designers capture work complexities and then can employ a variety of methods from experimental and quasi-experimental laboratory studies, naturalistic observation, computer simulation, to full-scale exercises, in order to better understand and support future technology and system design. Such problem/practice centered design approaches are at the heart of the CSE and human factors fields and as such, they can be exploited to support existing approaches to exercise and training design that already implicitly incorporate components of these problem classes. However, it is not often easy or readily apparent how to cope with the constantly moving target of task-artifact cycles in system design. This presents the difficult challenge of envisioned world design (Carroll, Kellogg, & Rosson, 1991; Dekker & Woods, 1999) which asks designers how do we design new systems when the introduction of new technology inevitably transforms the fundamental nature of practice (P. J. Smith et al., 1998). This is a challenge not only to the systems being created, but also the approaches and methods utilized to gain insight into those systems.

So far, this chapter has discussed how to successfully approach envisioning for staged and scaled world design from a CSE perspective. What is not explicitly covered in these approaches however, are the novel system challenges inherent in the planning of large-scale exercises brought on by new dimensions of scalability.

New Scalability Challenges

The ‘scaling’ process, in scaled worlds, assumes that researchers have already been looking at the world and have potential leverage points into the nature of practitioner
work. The prescriptive exercise planning process outlined earlier in this chapter can be seen as a stage for scaled world design complemented by these CSE techniques for discovery-a basic workflow that extends from planning how and what to run in an exercise, to developing rich challenging scenarios and incorporating passive observation in order to cultivate better learning.

To go back to the description of the response domain that began this chapter though, what is a unique challenge to traditional CSE designs are the new scaling dimensions introduced by these types of modern exercises. A number of scaling factors were discussed in the building block approach, factors that still must be considered, but this thesis looks at scaling from the perspective of a complex distributed cognitive system. The scaling dimensions defined here reflect similar emerging approaches for defining dimensions of macrocognition. These macrocognitive dimensions particularly characterize cognitive work in complex situations with regard to scope of consideration, temporal scales, and distributed interactions and task complexity (Patterson & Miller, 2008). From this perspective, this thesis defines three general dimensions of scaling. These scale factors do not just refer to the number of participants involved, or amount of resources collected, but rather they refer to the work and coordination difficulties that can arise over:

- **large temporal scales**: operations that last longer than typical one-hour sessions and can entail multiple hour, multiple day events

- **large spatial scales**: where supervisors and agents are spread physically over space with escalating independent and interdependent events occurring in multiple places and at multiple times
• **distributed agents**: coordination must occur at a distance as multiple parties coordinate and interactions between groups are no longer co-located and operate at different echelons.

Designing for scalability in a learning lab must consider these dimensions throughout exercise development and execution. Further scaling complications, such as those outlined in the building block section, can arise across the planning, staging, and execution phases of exercises. Among these are recruiting participants, synchronizing logistics, and soliciting the necessary time commitments over the different groups. Key personnel then become: the stakeholder groups who would commit resources to make the exercise possible; the different stakeholder groups who would benefit learning from the results; and the different groups who would participate in the actual exercise. Running such simulated events requires developing significant roles for all of the participating groups while keeping all groups meaningfully engaged even though unanticipated events and breakdowns occur. Another set of challenges complicated by scaling arises in the data collection and analysis process: how will observers track what is happening when events occur in different places? Given crises have a spatial extent and changing threats pressure different groups at different echelons engaged in the response process- people in different places will be observing different parts of the total process. Staging the exercise must prepare for a distributed yet coordinated observation process. Those who participate in crisis management exercises often learn a great deal, but how will these lessons be collected and combined to produce results that are valuable to others that beyond simply local and sharable learning as with the AARs? Developing this level of generalizable and sharable learning requires assembling
and analyzing the data to extract lessons for the different learning purposes tied to patterns of cognitive work.

These represent the motivating challenges toward designing a framework for conducting staged world exercises as learning labs. To explore how to overcome these challenges, this thesis next looks at how experts and professionals build the research paths that help uncover and explain those trends and cognitive themes. To accomplish this, three observations are outlined in the next chapter which serve as prime examples of multi-scale exercises that hit on a variety of these planning and scaling challenges that accompany any exercise. All of the key learning lab dimensions, stakeholder roles and goals, envisioning scaled worlds, learning elements, and scaling challenges are represented in these multi-scale observations. Through a meta-analysis of these observations, how this has been done in the field is reflected on. Being able to transitions from messy field operations in natural laboratories such as these operational settings grounded in authenticity, and successfully support the planning, execution, and analysis of scaled worlds leading to sharable learning and greater understanding is a significant challenge. Using this common class of criteria defined throughout this chapter, how these dimensions play out in real world operations can be seen and used as a guide to assess and stage exercises as learning laboratories.
CHAPTER 2

THREE LEARNING LABORATORY OBSERVATIONS

2.1 Overview and Rationale

The evidence from Chapter 1 indicates that current approaches to large-scale training exercise design may not be sufficient vehicles for learning on multiple levels. A lot of smart people have spent significant amounts of time and money thinking about and running these types of exercises, but the state of practice in complex domain, crisis response in particular, is relatively disjointed and the disaster response communities have experienced a number of growing pains attempting to scale up training efforts. The concepts underlying the learning lab seek to support these shortcomings and this starts by reviewing three real-world observations. The three observations described in this chapter reveal general patterns of cognitive work along a number of scaling and learning dimensions. Throughout this chapter, the goal is to expand upon the learning laboratory from a concept into a system for discovery. For now, the learning laboratory is an envisioned system; nobody has explicitly conducted one. However, a number of exercises of varying scales have been conducted that share the core CSE components of staged worlds, envisioning, and learning. This thesis proposes that
these properties and their support functions form the basis of a framework for the learning laboratory approach.

To set up the framework, two analytical conventions will be used throughout these examples. First, a variety of process tracing techniques (Woods, 1993) are utilized to illustrate multiple perspectives in different contexts throughout each exercise. In establishing this learning laboratory criteria, each process tracing figure represents varying situated functional representations that help to maintain multiple levels of target:mapping traceability in each exercise. Second, brief overview tables (similar to the one in Table 2.1) appear at the end of each exercise highlighting specific instances of key learning lab events as laid out in Chapter 1. Any exercise that shares these basic dimensions can described as a learning laboratory.

<table>
<thead>
<tr>
<th>Stakeholders/ Stance</th>
<th>Envisioned World Attributes</th>
<th>Scaling-challenges and exploits</th>
<th>Learning-challenges and exploits</th>
<th>Staged World Components</th>
</tr>
</thead>
<tbody>
<tr>
<td>Multiple stakeholders &amp; practitioners</td>
<td>Balancing across Plurality/ Underspecification/ Groundedness/ Calibration</td>
<td>Distributed groups Spatial Scales Temporal Scales</td>
<td>Local, sharable, generalizable, learning Usability, usefulness, understanding</td>
<td>Scenarios, probes, pacers, blocking, rally points, forums for reflection</td>
</tr>
</tbody>
</table>

Table 2.1: The five basic dimensions underlying a conceptual learning laboratory.

The first half of this chapter provides a detailed meta-analysis of three field observation studies as partial instances of learning laboratories. To frame these three exercises, the goal is to examine how these exercises instantiate learning lab components and review how they relate to scaling, learning, and resilience dimensions.
from an observation and analysis level. The second half of the chapter synthesizes these findings into an initial learning laboratory framework to first guide a series of interviews in Chapter 3, and then in Chapter 4 develop a planning tool design seed that guides organizers across the planning, observation, and analysis stages of an exercise.

2.2 Stance, Limitations, and Scope

From a CSE perspective, we are typically concerned with the design of new systems. People conduct these exercises for a variety of reasons and some have to do with addressing better training. This thesis addresses a very broad spectrum of cognitive work beyond skills training to explore how approaches to cognitive training can translate into new systems. This sets up a basic dimension for the learning lab whose goal is to translate from local learning to generalizable learning. Often when we as CSE researchers write up studies, there's a tendency to skip over the process and only talk about the outcomes. For example, if one looks at a typical CSE paper and a typical experimental psychology paper - with the latter, the reader knows exactly why the authors are making their claims. Their stimulus world and their experimental design are so constrained that explicit methodology and tractability are usually quite formal. In such studies it is easy to trace what piece of data led to what specific theoretical conclusions. This formalism tends to get a little muddier in CSE given the exponential complexity that accompanies the real world. It is rare to be able to track how specific pieces of data led to specific theoretical conclusions. To study the world, we shape the conditions of observation through the design of scenarios and problems, however, researchers tend to then cite only the theoretical conclusions leaving out the methods.
and processes that lead to discovery. There is the risk that the mechanisms of how those things were built is lost when more interest is focused on the general themes. To that degree, in the following three observations the data will be described in detail so as to be able to form an abstract discussion of what it means to do CSE. Given the difficulties of conducting exercises, the goal is to capture some of that difficulty along the specific learning lab dimensions to reflect on how those difficulties manifest and can impact learning opportunities in a variety of ways.

In each of the exercises below, CSE practitioners came in fulfilling observation roles which limits the degree to which we can directly relate how these findings would fulfill the defined learning lab goals. The CSE individuals were not sufficiently weighty stakeholders in these exercises and therefore they did not get to define the roles and purposes in the exercises (to complement this, interviews in Chapter 3 focus on exercises where CSE individuals were primary stakeholders). In many real world scenarios, the CSE person is almost always a team member, but not always a team leader. It is often rare that CSE individuals get to design training exercises from the start. Because of this, part of the discussion in each case below addresses envisioning alternative opportunities to consider as if one did have total control from top to bottom. Because of this ‘piggy-backing’ on top of found case exercises, it is very important to describe the stakeholder goals responsible for planning in each of the exercises up front. For example, a trainer designing an exercise often has the goal to see if particular individuals can do their job. They are not necessarily designing it to learn how this job should be done or how it could be improved. A trainer designed exercise is like a teacher giving a test- they are not designed to be opportunities for exploration or revealing about the area of study but rather they are designed to evaluate an
individual’s knowledge. At the other end of the spectrum, we often hear of exercises designed for success that not really very challenging. The purpose of many Army exercise designs, for example, is not how to make a better army per se, but to evaluate whether those individuals are fit to go into battle. Both of these ends represent a huge disconnect that must be acknowledged and understood before being able to abstract higher level evaluation and assessment from an exercise. Even in the best case, there is a high upper-bound of what you can learn. Across each of these three exercises, we have to acknowledge there were limitations and given the established plans and goals of the stakeholders, there is a limit on what we as researchers can abstract that is directly related to why each event was designed. To this degree, we need to still be able to maximize learning effort presupposing if someone else did the planning, or even worse, if someone else had done the planning and the analysis. Maximizing learning yield therefore depends on being able to generate these goals, familiarity, and a map of events in CSE terms.

The analytical conclusions arrived at in this chapter are the result of participation in a large variety of field exercises, operational systems, and exercise planning sessions over the past six years across multiple domains. Three exercises in particular stand out as rich experiences that can be described as learning labs from multiple scales. The first observation was a found case, a military intelligence observation at Fort Huachuca. Here our research team saw unique pedagogical approaches toward training analytic problem solving in a linear war-game simulation set in WWII Great Britain. The next exercise was a future incident training event conducted by the FDNY. This urban fire fighting joint emergency operation, titled Kings Mall, was observed in both tabletop form and full scale execution. This served as a rich opportunity to observe the results of
a detailed planning process and training techniques first-hand and see how stakeholder goals evolved from planning to the run-time exercise. Last, a large scale technology demonstration, Strong Angel III (SA3), illustrated the significant challenges to both envisioning and learning in large scale exercise design. As a participant and observer in this simulated large scale biological incident demonstration, clumsy technology coupled with brittle exercise planning was seen leading to an escalating irrecoverable exercise breakdown. In addition to these experiences, through participation in a number of tabletop simulations, robotic search and rescue technology demos, DHS course-work, and field-work with a canine search and rescue team, there are many valuable lessons to be learned across all of these that are applicable to the design of learning laboratories.

In describing each of these observations, analysis is aimed at identifying how multidimensional scaling challenges impact learning and understanding. Run-time exploits and difficulties alike played out in all of these exercises that point to general challenges in design and observation that led to under-utilized learning opportunities. General goals and stakeholder purposes are defined for each exercise to relate why they were created followed by how they were conducted. Surprises and challenges emerged in each of these events which will be discussed along with key leverage points and lessons learned. Implications on methodological techniques and future learning challenges will also be discussed. To describe each of these field exercises as containing elements of learning laboratories, a basic abstraction hierarchy is used to outline why we studied what we did, what we were studying, and how we did it for each case. Across all of the observations, the top level goal for us as researchers was to uncover the challenges to practitioners at work. We have learned a significant amount from
these specific studies with regard to understanding and sharing these research results. Analysis and research from these specific observations has been reported in the CSE and human factors literature covering such diverse patterns as anomaly response, data overload and information synthesis, clumsy automation, coming up to speed transitions, and common ground challenges. Here, the goal is not to simply restate the same findings and relate the same outcomes from the studies, but rather the goal is to relate the processes and challenges faced that led to those discoveries and convey what it means to organize and plan for learning from large scale exercises.

2.3 Ft. Huachuca

2.3.1 Overview and Scaling Dimensions

The explicit goal of the Ft. Huachuca exercise was to train recently promoted US Army military intelligence Captains on a new intelligence analysis technique. A multinational instructor team was responsible for training 40 individuals a number of analytical approaches, with added emphasis on a new form of analysis, Analysis of Competing Hypotheses [Heuer, 1999]. As a learning lab, our observation team was able to ultimately develop a good understanding of the cognitive challenges of intelligence analysis as well as deconstruct the training exercise in real time to identify how pre-built and ad-hoc elements of the exercise served as key staged world components. This led to a large amount of research that serves as pointers toward identifying how one could do intelligence analysis better. This rich understanding was only possible due to the resilient observation system we built.

We built a significant amount of observer flexibility into the observation and this greatly contributed to the capabilities of the observation team. The original
exercise plan arranged by the training team, for logistical reasons, was to have four simultaneous sessions occurring at the same time. As CSE researchers, when we learned of this upon arriving, we saw an advantageous situation that we could exploit. The trainers had organized the exercise this way for efficiency and expedience for the commander, however, we saw this as a chance to observe four similar groups do the same exercise simultaneously. Ultimately, this made a more resilient observation system because observers were able to share events in real time. For example, on multiple occasions, one observer knew the commander was on his way to a particular room and by knowing what had happened in a previous exchange ahead of time, the observer could focus on key similarities and differences in their own upcoming session. This made the observation more robust because of these anticipatory cues (Nysser & Javaux, 1996). Observers, in discussing with each other, could actively discuss and cross-cue one another what to look for—almost a form of collaborative cross-checking for observation (Patterson, Woods, Cook, & Render, 2007a). If we had conducted the observation in a more traditional experimentalist approach, we would have isolated each observer, they would have no communication at all, and then they would bring all their observations together and validate them against one another at the end of the exercise. Such repeated measures type approaches would be less resilient and illustrates a key difference between CSE methods and traditional approaches. Instead, we used these four instances as a self-regulating way of building resilience into our observational system. Being able to simultaneously learn about the event itself and the different ways the captains reacted to the same information and events with literally the same challenges, the observers were able to capture a rich level of cognitive work adaptations. Again, because we did not have any control in the design, we could not
interject or use think aloud protocols; all observations were opportunistic and not invasive. By cross coordinating, observers could compare and contrast team differences in real time about aspects of each scenario. This proved quite interesting given that separately observed events played out slightly differently in each team. Given that the linear structure of the scenario depended on specific interim outcomes, it was valuable to capture the different approaches and analytical processes undertaken by each team to get to successive scenario points. The cross-cueing large scale observation enabled much deeper observation around these points that might otherwise not have happened. This simultaneous observation was initially an unplanned event and adaptively we were able to use it in a clever way to achieve a high level of observational agility. Overall, the elements that made Huachuca interesting from a learning lab perspective are:

- Four distributed simultaneous observation sessions
- Found Case Training Scenario
- Building familiarity by participating before observing
- Developed a scalable resilient approach to observation

2.3.2 Exercise Details

In March of 2005, CSEL assembled an interdisciplinary group of cognitive engineers, political scientists, psychologists, and intelligence analysts to participate in and observe a high fidelity training scenario being conducted at the U.S. Army Intelligence Center at Fort Huachuca, Arizona. Fort Huachuca is the training ground for all U.S. Army intelligence analysts. This work has contributed to a significant amount of our own
research base in intelligence analysis findings and has been reported in a number of previous studies (Trent, Voshell, Grossman, et al. 2007; Trent, Voshell, & Patterson 2007; Trent, Woods, & Patterson 2007). The previous work has described a number of cognitive challenges intelligence analysts face including cognitive work balance tradeoffs, uncertainty under data overload, creeping validity in analysis, as well as the multiple methodological components of the study. What has not been thoroughly reported, however, are the challenges and methodological techniques our team used in observing four distributed teams simultaneously over the course of one week. This section outlines the valuable learning experiences from the observation and their implications for large scale exercise design. The most valuable of these lessons learned was that in order to stage these large scale events effectively, one has to plan how to coordinate multiple observers.

Ft. Huachuca serves as a good lead-up to large scale exercise design because it gave our team a way to test how well our observation techniques could handle large scale events. Also, Ft. Huachuca matters because it shares multiple dimensions of a large scale exercise. On a temporal dimension of scale, it was a five day exercise where for eight hours a day, multiple teams of observers monitored distributed groups of analysts while a roving instructor groups ran multiple locations simultaneously. Traditionally we would have observed and run one group at a time, however, the simultaneous execution gave us the equivalent ability to test how well we could conduct a distributed observation of the multiple teams. Although technically these were not truly distributed interdependent teams, the groups simultaneously ran the same exercise independently. Our spatially distributed observation team had multiple challenges to overcome. What became interesting was how the scenarios played out
slightly differently across each group as well as how instructors shared in monitoring and intervention planning. Since events unfolded simultaneously in real time across the same general space, observers had the opportunity to test out how to monitor, collect, and analyze data in what was tantamount to a large scale event.

The strength of the study lay in the fact that for each analyst team in the exercise, we had four groups of individuals that were pretty much identical in their backgrounds. They were all taking the same class and training for the same military positions. As military intelligence captains, they are analysts and analyst supervisors with a bachelors level education and an average of four years of military service. On this occasion, the class of 40 captains was randomly divided into four independent squads for the fictional counter-insurgency training scenario based in World War II Great Britain. In the scenario, students simulated the role of a German S2 shop coping with a growing insurgency. Instructors played the role of unit commanders responsible for pacing the flow of new materials and providing feedback to the students. Observing each squad proved to be an interesting scaling challenge for us as observers on two fronts. First, for ethnographic observation in general, when utilizing observers and researchers not intimately familiar with a work domain in advance, one runs the risk of overlooking critical vulnerabilities- the general challenge posed by the law of fluency (Woods & Hollnagel [2006]). We needed to stage an observation that could scale up the observation teams expertise. Coupled with this challenge was one of spatial scaling, we needed to also be able to observe multiple-group interactions and simultaneous processes taking place in multiple areas in parallel.

To address the first issue and to familiarize ourselves with the domain, the scenario, operational semantics, and the anticipated work of the practitioners, the observation
team took part in a two-day accelerated version of the exercise prior to observation, enabling us to face the challenges first-hand prior to the exercise. After this learning experience, significant discussion time was spent with the instructors reviewing the details of the scenario, instructional objectives, and their own personal thoughts. From this, the research group amended the original list of challenges they had assembled prior to the exercise and discussed these in detail so that everyone was on the same page before observation was to start the next day. To address the second distributed observation challenge, the team developed a unique variation on established ethnographic techniques to conduct a detailed observation of the multiple groups of analysts. The goal was to embed a single observer in each room and collect as much detail as possible, ramping up resources when necessary, and use collaborative instant messaging software to perform a distributed protocol analysis in real time across observers between the multiple analyst squads working independently on parallel tasks. A meta-observer group was set-up in an observation command center which coordinated and tracked the multiple problem solving processes and comparative progress while the analyst squads faced the escalating (simulated) counter-insurgency, allowing a big picture view of each teams’ work to emerge. This provided researchers not only with a robust study, but enabled the deliberate staging of the conditions for observation as those of an analysis problem in its own right. With all the analyst groups performing in parallel, the observation team was in the unique position to use the protocol analysis itself as a form of information analysis, just as the analyst teams were doing.

From a learning perspective, intelligence analysts experience many cognitive challenges that extend beyond the barriers of individual problem-solving. The exercise took place in an educational environment where instructors were not merely training skills
for practice, but rather, it functioned to provide the basis for further skill acquisition (Trent, 2007). From the research perspective, the learning goals the research team identified coming out of the study were to build on earlier conclusions about the nature of analytical work and the cognitive challenges that it presents (Elm et al., 2005; Trent, Woods, & Patterson, 2007), reflect on challenges to team cognition, as well as evaluate teaching effectiveness. Not only did this share similar patterns of cognitive challenges with other domains, but being an educational setting, the exercise was fundamentally structured around facilitating successful local learning. Instructors faced the challenge of maximizing learning in both an efficient and timely manner. This pool of instructors responsible for teaching and preparing the students for the problem solving necessary in an operational theater can be seen as one set of stakeholders. The students, as participants, were also a stakeholder group vested in the sense that their local learning will contribute toward furthering their careers as well as prepare them for operational experience. Last, the researchers were a stakeholder group interested in the cognitive challenges in the work domain so as to be able to abstract and find general patterns of work.

2.3.3 Designing For Coordinated Observational Agility

The scenario was a found case which the researchers did not have control over designing. With found cases, rather than designed problems, it is critical for observers to be able to prepare to be surprised and be able to recognize what groups were doing in what contexts so as to be able to recognize when behavior deviated from typical. When designing staged worlds from scratch, this is accomplished by creating probe events and observing how individuals respond organically within the scenario.
Figure 2.1: Five CSE Research observers were able to simultaneously watch the same events while communicating their observations with each other, and with a command center. Instructors watched and interacted with each analyst team individually, while researchers observed the interactions across all instructor groups and within each team. For each analyst team, the scenario and challenges faced played out slightly differently at run-time.

In the case of this exercise, we were able to rapidly familiarize the team with the domain by participating in the exercise before observing it and then re-calibrate our observation plan to identify components of the scenario that functioned as knowledge probes ahead of time. The research plan was to create a distributed observation system with forward observers in each team talking back to a command post, as well as with each other, to keep track of problem solving processes while analyst teams worked the counter-insurgency information analysis. Prior to the exercise beginning, all analysts teams were provided an overview briefing by the instructors that described the general situation. Teams were assigned to separate rooms, each equipped with dry
erase boards, maps, overlay material and four computers with basic office productivity applications. Each team received a pile (literally) of information representing the remnants of information from the last intelligence section (S-2) which had been recently destroyed. This is a common scenario element CSE often uses in staged world exercises-starting agents in the middle of the action. Teams were given an initial period of four hours to sort through their information, after which they received a new report, or serial approximately every two hours. These reports served as pacers (Schoenwald, Trent, Tittle, & Woods, 2005) that kept the scenario moving. Similarly, if groups faced roadblocks or simply did not make progress, instructors could intervene and steer students back on course. These instructor interventions both served as impromptu team reflections as well as helped to keep the pace of the exercise moving. Because the exercise was so long, multiple hour fixations and interruptions did not compromise the entire event and could be dealt with by instructors effectively throughout the exercise run-time.

Analyst teams had to analyze the available data and task resources for further collection based on their assessments. Many of these assessments required direct actions communicated from the teams in the form of requests for information (RFIs). All RFIs were collected and served as key artifacts around which to trace team processes. At the end of each day, teams briefed their assessments to an instructor acting as the commander. These command briefings served as valuable points of reflection, for both the instructors gauging team progress, as well as for observers getting access to externalized cognitive work that could be compared against the day’s observation. Each serial was designed to provide realistic reports of activity representative of a counter-insurgency in 1940s Great Britain. Because each report had certain normative
interpretations and multiple other suboptimal interpretations, they also served as embedded event probes which would elicit various analytical vulnerabilities. Since all teams received the same probes at the same time, they served as opportunities to observe repeated measures between groups. By having experienced these in advance in the warm-up, the observation team used the probes as landmarks in an event map for each day’s observation.

In order to observe all squads concurrently, one observer was assigned to each squad. Utilizing one constant observer with each squad throughout the course of the exercise allowed rapport to be built and fluid communication with the analyst teams. The observer was closely integrated in each analysis squad and this enabled more natural behavior to emerge throughout the course of the exercise. Two other observers formed a meta-observation group and established a command center in an adjacent room from which to monitor and facilitate communication between situated observers. The primary investigator retained the ability to move between squads and the command post. Similarly, an ad-hoc wireless network was deployed enabling communication between observers and the PI and the command post. Each observer used a laptop (Windows and Macintosh platforms) to communicate with the command post over a wireless 802.11x connection. Observers took notes by having conversations with the central command post using asynchronous instant messaging software (all utilizing the ZeroConfig protocol standard).

Inside the command center two to three members of the observation team monitored each stream of incoming observer data. The command observers could easily perform cross-checks across the situated observers to detect emerging trends or direct attention
toward anticipated events. The combination of synchronous and asynchronous communication tools provided status-at-a-glance cues that helped to dynamically calibrate focus and assisted the observation team in coordinating a near realtime distributed protocol analysis. The command center could effectively coordinate and redirect resources and observer attention constantly and at specific rich probe points. At the end of each day full transcripts of each squad’s time stamped chat logs were combined together with any supplementary multimedia material. Each of the chat programs allowed observers to transfer files as well as initiate one-way and two-way audio and video connections with the command post, allowing for a very robust time-stamped data collection as well as providing the command post with a direct remote presence.
2.3.4 Lessons Learned

The one big lesson learned from Huachuca was the power of coordination amongst observers simultaneously observing different teams performing the same work. Methodologically, this was unique and it is one way where the approach differs tremendously from a traditional experimental design. Emphasizing and supporting the necessity of observation is a staple of staged world exercises and by proxy, a fundamental
component to anticipate and build for in a learning lab. The learning in Huachuca represented accounts of observational adaptations for studying distributed work in context and revealed profound challenges in both the cognitive work of intelligence analysis, as well as the challenges of conducting observational studies at an increasing scale. The observation itself was just as much an example of distributed work as the analyst activity requiring intricate coordination to work together, prepare assessments, and brief each other at different intervals. This approach was accompanied by rich and varied data collection and analysis looking at the interplay between teams of observers as analysts, observers watching other observers, and observers watching teams of analysts do distributed work. By making these challenges explicit, they could be dealt with as they arose and incorporated into future training exercises and distributed observations. Overall, this observation with multiple levels of communication in a networked observation team is a more resilient approach to observation.

With regard to learning about how to conduct instructional exercises better, Grossman et al. (2007) looked specifically at instructor interventions in the exercise. Similar to observer controllers, instructors staged various interventions through the week long exercise. The interventions represented both the goals of the training exercise as well as the instructors’ assessments of group activity. As Grossman et al. (2007) point out, these were also valuable in that they illustrate how expertise is passed in an organization. Interventions were discrete events throughout the day and also culminated at the end of each day in the form of commander briefings. With the explicit goal of doctrinal teaching (of the ACH method) instructors also served to impart broadening checks (Woods & Hollnagel, 2006) in cadet processes to counter a common cognitive challenge in intelligence analysis and synthesis, premature
narrowing. They also actively worked to keep the teams engaged in the scenario. While [Grossman et al. (2007)] found that interventions were not always effective, the study does illustrate the demand for such functional roles in a learning laboratory. Instructors did not have access to a meta group responsible for fulfilling the research base and learning goals, which would enable instructors to concentrate on observing and engaging the teams. By the end of the observation however, the instructors began to enjoy exploiting our own observation resources to quickly get better insight into team activity before rotating in to a session.

Figure 2.4: The annotated timeline for one of the analyst teams depicts the relationship between observers, stakeholders, and practitioners against the events of the Northern Star scenario.

The second lesson learned regarding this distributed observation team was the value of the meta group opportunistically providing top-down observer coordination support to richly observe and relate the concurrent, distributed work of each team in the exercise simultaneously ([Trent, Voshell, & Patterson, 2007]). The observer command post meta group supported and grounded the distributed observers. The command post provided a global frame of reference for the observation as well as a
domain expert on demand for the scenario who functioned to help cross-cue observers in real time to detect emergent trends and direct attention to instructor actions and anticipated events. This helped the observers notice what was interesting ‘on the fly’. The entire observation team benefited from being able to dynamically calibrate, focus, and perform a near real-time distributed protocol analysis. As mentioned, by the end of the exercise the instructors themselves utilized the ability to passively monitor team activity and deliver more effective interventions. This observational agility helped the observations teams essentially understand a problem map to be able to identify inherent difficulties and challenges, based on the learned vulnerabilities experienced first-hand as well as knowledge elicitation with the veteran instructors. By realizing how difficulties were built in to the exercise, a much higher level understanding about the nature of the exercise could be formed beyond what could be gained from, for example, a simple master list of events (i.e. the MSEL). For this to work as successfully as it did, the observers had to understand that they had basically three things they were doing at any given time: communicating with the command post, communicating with the other observers, and their own embedded observation. It was critical that all observers understood that the purpose of the first two was to support the third one. The resulting high resolution of observation coupled with constant high level commentary served as a rich basis to create process traces from. As mentioned in the first section of this chapter, analysis for a learning laboratory exercise advocates detailed multi-perspective process traces throughout events. Examples showing the abstraction and refinement of different aspects of these in Huachuca are highlighted in Figures 2.4 and 2.3 to convey how the raw data was abstracted and rendered in a format to analyze at a higher level. Furthermore, Table 2.3.4 following this
section represents an overview of Huachuca highlighting the multiple scales and CSE
dimensions that are characteristic of a learning lab approach.

The Ft. Huachuca study served as a great test of observational agility and represents
a broader scale exercise that showed that our observation team could conduct a
relatively large scale observation effectively and affordably using basic off-the-shelf
hardware and software. In a sense, there was a degree of envisioning to the scenario.
Even though it was conducted based on quasi-historical events with no technology, the
hypothesized demands that set up the scenario were based on competencies the school
envisioned analysts would soon be facing as they transitioned to an actual theater
of operation. From the cognitive demands and challenges we discovered, we were
able to eventually revisit instructors and share these as insights for future training.
The research was published in various avenues and we were also able to conduct a
workshop based on the findings. By further sharing this general learning both within
the domain as well as other communities outside this field, the goal was to stimulate
the innovation process for both future technology support designs, as well as a means
to familiarize other researchers with the more abstract challenges to understanding
collaborative analytical work.
Table 2.2: Huachuca as learning lab element overview. Instances of observation are marked underneath the five learning lab concepts.
2.4 Kings Mall- Urban Firefighting Exercise

“Every 10 years another term comes out. In the 60s, ‘copious amounts of water’. In the 70s, ‘synergistic effect’. Now it is ‘situational awareness’. What is the only element that will be present at every situation I show up? The one element- certain that will exist on every response, is me. The keys to the kingdom and secrets of the universe are personal development. Looking at the intuitive and recognition primed stuff, it all means we need to become very well versed- a wise man learns from his own experiences, a wiser man learns from the experiences from others. You need to train.”

- Chief Robert Burns, personal interview, 2008

2.4.1 Overview and Scaling Dimensions

The Kings Mall exercise represents a different training emphasis than Huachuca and brings us right into a modern crisis response case. Given recent and anticipated near future concerns, the training goals for this urban firefighting exercise were specifically based on creating new cognitive skills for incident commanders. Efforts across the fire department are being employed in an attempt to create new real-time problem solving skills to help incident commanders learn how to assess their response actions in unknown situations. This is interesting because these exercises are typically quite skill based- a procedure is taught and then one needs to prove they know how to do it. Here however, they are facing the difficult challenge that it is not very easy to teach someone to be surprised. While training schools and multiple third party contractors offer handbooks and checklists, i.e. “when confronted with a new thing, recognize it’s new, and do the right thing”, we knew from the start there was a CSE leverage point for support here. After a significant amount of our own doctrinal review of the fire department combined with real world experience on response operations, we began to see that these basic training approaches left them in the same shape as Douglas Haig on the Somme (Shattuck, 1995). As part of a larger CTA, we wanted to see how we
could better help the fire department prepare to be surprised, link analysis to expertise, and empower them to be better advocates when confronted by new technologies.

Kings Mall was a found case exercise developed by an internal fire department training team that had been looking to scenario based design approaches as one way to approach training that overcame these traditional shortcomings. Like in Huachuca, there was no CSE input in its initial design, however, we were able to observe the exercise from its planning to its execution in order to assess the evolution of desired learning goals and their outcome. For the full scale exercise itself, from a learning lab perspective, the CSE observation team was able to implement a similar observation technique as in Huachuca to analyze multiple distributed teams simultaneously coordinate in a response. Second, this exercise was intended to not be purely a training exercise. This intent, however, revealed a number of competing institutional goals in pulling off such an endeavor. For people to learn, many would argue that they have to face failure conditions to learn from. Firehouses choose to participate under the auspices of training, but similar to not letting pilots crash the simulator, cultural norms prevent these failure conditions from truly occurring. With this in mind, the elements that make Kings Mall interesting from a learning lab perspective are:

- Training goals centered on learning and cognitive training
- Distributed coordinated observation over a (large) temporal scale
- Multiple distributed groups conducting a coordinated response
- Coping with cultural norms of the domain
Crisis management as an overall system is in a state of flux. For the urban firefighting community, events such as the 2001 terrorist events in New York City and the American Airlines Flight 587 crash in Belle Harbor were surprise events that became harbingers of organizational change. Fire departments, especially those in critical commercial and cultural centers, have been re-focusing organizational efforts to support strategic, tactical, and operational response and training across three core dimensions (Voshell et al., 2008):

1. A changing threat environment,

2. New demands driving the need for technology adaptation

3. Growing shifts in organizational contexts for expertise

Chief Joseph Pfeifer, stresses that leaders must move beyond traditional reactive behavior by demanding resilient and adaptive approaches for managing complex incidents at both strategic and operational levels (Pfeifer, 2005). To this end, these larger scale events challenged the organization. One particular response to these three diagnoses has been the increased use of future oriented training exercises. Under the direction of Chief Pfeifer and Chief Neil Hintze at FDNY’s Center for Counter Terrorism and Disaster Preparedness, their group has been staging an increased number of exercises with the objective of cultivating learning to prepare decisions makers for operations in these new environments through large scale joint training exercises. As part of a multi-year ongoing cognitive task and work analysis, we were able to assemble a research team to participate as observers in the conduct of one such large scale training exercise at the Kings Mall Shopping Plaza. Kings Mall was framed as a future incident case and while we did not get to take part in the actual planning of
the exercise, we observed the exercise as both a planning tabletop and then as a full scale execution. Being able to see the entire life cycle of the exercise enabled us to see: how the results of the planning process propagated; appreciate what a great deal of effort it takes to pull these sorts of events off; and identify exercise implications on local, sharable, and generalizable learning.

In July, 2007, as part of a joint effort between CSEL and the United States Military Academy at West Point we had the opportunity to observe a large scale joint training exercise conducted by FDNYs Center for Terrorism and Disaster Preparedness. Based on a prolific amount of interviews, knowledge elicitation sessions, and operational observations as part of a CTA of urban firefighting, a significant amount of insight was gained about the Fire Department as a learning organization. Different themes from this CTA have been published in a variety of formats (Voshell et al., 2008; Trent, Voshell, Fern, & Stephens, 2008; Fern, Trent, & Voshell, 2008; Peffer et al., 2008; Branlat, Fern, & Voshell, 2009). Similar to Huachuca, the goal of this review is to focus on the mechanics behind how we arrived at specific learning lab research themes. The Fire Department is a highly vested stakeholder across all operations, simulated and real. This goes hand in hand with the driving demands it faces to maximize learning across a number of core dimensions. Fire departments currently face a changing threat environment as well as dramatic reductions in experienced practitioner staffing (Fern et al., 2008), and these constitute key stressors for organization resilience. For example, after 343 firefighter lives were lost in the September 11 attacks, 2917 firefighters and officers retired over the next year representing some of the most experienced across the department. This created a large experience vacuum with approximately 60 percent of new hires fresh to the force (Griffith, 2005). Although these numbers are unique
to FDNY, catastrophic events can create similar expertise deficits in other high risk organizations. This work provided an interesting opportunity because unlike some domains that CSE studies, the fire department maintains a committed demand across many aspects of work culture for constantly leveraging its own learning capabilities.

2.4.2 Kings Mall From Tabletop to Boots on the Ground

The Center for Terrorism and Disaster Response Preparedness has established a name for itself advocating scenario based training as the way to address the changing dimensions faced by modern urban crisis responders (Hintze, 2008). The Center is constantly evaluating different approaches to scenario design for training as well as different teaching methodologies. Philosophically, their full-scale exercise approach is primarily geared to enhance familiarity between units. For this exercise discussed below, this emphasis was particularly aimed at improving EMT and Fire coordination with a secondary goal of familiarizing personnel with new physical locations. The exercise followed a similar planning schedule as prescribed by FEMA/DHS guidelines and was preceded by a table-top exercise to gain a general, high level overview of how the full-scale exercise would unfold. The tabletop planning exercise lasted three hours and incorporated multiple stakeholders including senior fire and EMS personnel, the fire marshal’s office, as well as local mall employees. The table-top first consisted of a familiarization process where senior officers asked the mall staff representatives a variety of questions relating to location and resource attributes. Additionally, fire personnel were able to familiarize themselves with the malls existing emergency procedures. From the table-top observation, the planners were mostly concerned with manipulating time constraints (i.e. having the building be near the edge of a borough)
Figure 2.5: Annotated exercise staging overview relaying how planning goals led to events in the tabletop scenario, which were then complemented by additional events during exercise run-time.

as well as using the scenario to better acquaint inter-agency fire, HAZMAT, and EMS needs and responsibilities with outside organizations including the police and mall staff. Strategically, the goal of the tabletop was to familiarize officers and collaborating organizations with what constitutes critical information during a response. The second part of the tabletop was a general walkthrough of the scenario events. Chiefs and officers talked aloud what they were considering and the basic escalating series of events were run through. While the table-top did provide some insight into the expected flow of work practice (Flach, 2000), it also provided the observation team with a grounding in the planners’ performance objectives (much like running the exercise prior to observation in Huachuca).

One month after the table-top, the Kings Mall scenario was conducted as a full-scale joint training exercise for interagency planning and coordination. The observed
scenario simulated a mass casualty terrorist event that took place early morning, between the hours of 0215 and 0327, at an urban shopping mall. Such events are not often experienced or trained for in the US. The scenario was designed to place novel demands on rescue and response strategy and tactics. Additional elements were introduced to imply serious safety and security challenges to the responders themselves. This large scale exercise included multiple fire battalions collaborating with specialized units, emergency medical personnel, and police. Approximately 200 responders interacted with 50 injured confederates. According to our definition of large scale, multiple distributed groups had to coordinate and a significant amount of resources and personnel were involved across a large physical area. While at first glance the temporal duration seems relatively short, especially compared to Huachuca and the Strong Angel exercise described next, for this domain very large urban fires rarely exceed an hour in duration and this was of a similar time pressure to such large responses.

2.4.3 Observation and Analysis

Much like in Huachuca, being able to capture multiple points of view around situations in parallel was a key strategy in planning for this observation (Trent, Voshell, Grossman, et al. 2007). Four observers were pre-positioned in critical information-rich areas on the exercise grounds based on three general patterns of cognitive work using an edge-centered testing approach (Rousseau, Easter, Elm, & Potter 2005). One observer was situated at a rally-point between tactical fire operations and medical personnel, a key area to observe inter-agency coordination hand-overs and potential breakdowns. The second observer was placed within the incident command post to
monitor command and control ‘scaling-up’ activity (Oomes & Neef, 2005). The two remaining observers shadowed the observer controllers (OCs) who were responsible for driving the scenario in situ. The observation team monitored all tactical and command communication channels and documented the event with audio, video, image capture, and notepads. The observation team coordinated with each other via radio, phone, and short face-to-face discussions. This enhanced the ability for observers to cross-cue teammates to anticipate times of particular interest and heightened activity. For example, at one point an OC wished to further challenge a unit and spontaneously introduced a missing civilian inject. Observers cross-cued this information to one another and were able to concentrate resources to monitor the reverberations of the key event across the tightly coupled but distributed responders including the officer’s unit, command response, and the remote F.A.S.T. truck responsible for unit rescue. The following day after the exercise a quick process trace was assembled and we had the opportunity to interview the exercise planner, who was also the lead OC, to reflect on the exercise directly.

Again, like with Huachuca, this observation led to a significant amount of raw observational data to analyze. Observers constructed an initial timeline and marked the multiple probes that had been formed in the exercise. From this domain level timeline, we were able to start constructing process traces at varying levels of abstraction. Multiple functional analyses of urban firefighting were developed (Fern et al., 2008; Peffer et al., 2008) which served as tentative models of work and then used to test against the process trace as we started to formulate more cognitive descriptions of the domain. In terms of facilitating generalizable and sharable learning, these resulting functional analyses (as seen in 2.8) served as collaborative artifacts used in later
Figure 2.6: Detailed Process Trace analyses from the Kings Mall observation.
discussions with stakeholders which served to expose more detail around the goals and sub-goals that manifest in urban operations. These also served as the design requirements bases for new proposed technology systems (Trent et al., 2008).

2.4.4 Lessons Learned- Scaling at the Right Levels

While many coordination breakdowns were observed in the response, such as between EMS and rescue as well as between HAZMAT and most other groups, what seemed to be underutilized was the intent of testing the higher level command and control decision-making of the chiefs and incident commanders. This goal was present in the table-top and one the key areas of concern the department faces. The physical scaling of the exercise and the incorporation of all hands and all resources did not correspond with the virtual scaling that stakeholders desired to evaluate in this case. The stakeholders and planning team worked together, however, as they scaled the exercise up, planning goals specifically for the incident commanders were compromised for the sake of catering to the large number of individuals on scene. Skill training of individual units eclipsed the opportunity for any cognitive training of the ICs. The notion of a learning lab, emphasizes that stakeholders must work together to establish exactly what they want to train, there has to be a tacit agreement toward goal alignment. In order for those higher levels to get trained, larger contexts must be created. Based on a series of personal interviews with military instructors and exercise participants, interviewees all acknowledged a fundamental lesson learned from military training regarding scaling. They saw that what a lot of organizations do wrong is that they bring in all the physical components of the organizational hierarchy along with
them to the exercise. This can be quite brittle and training goals aimed at different contexts can quickly conflict and undermine one another.

At Kings Mall, as soon as dispatch gave the alarm call, all the resources started to arrive. As related in the process traces, firefighters were soon hauling gear around, setting up the conditions to charge hoses, and practicing everything from search and rescue, to laying hose, to specialized rope extractions under the auspices of only two prime observer controllers. Throughout all of this, the incident commanders were never really exercised because the pace of the exercise was literally driven by how fast ground units were moving. The ICs quickly became OC support staff. Just as much, all the chiefs knew the canonical path of exercise events ahead of time. The mall exercise became unit focused and in so much it was pressuring coordination in the face of operational tactics. Units were responsible solely for their own tasking and they were not looking for surprises (this was evident in all teams missing the IED bomb probe in their inspections). While OCs did an impressive job balancing catering observation and catering injects to individual units, they did not have the active ability to monitor and track these higher level learning objectives at the same time. The hot-wash following the event touched on some of these shortcomings but overall, given the scale and resources expended, many others could have effectively benefitted.

The nature of the domain itself adds further physical scaling issues. As large and as distributed as the physical operation was, it would be difficult to get much larger. Real world constraints start to impact this given that organizers cannot set buildings on fire, they cannot shut down multiple city blocks, and they cannot take hundreds of active duty personnel off duty. This would serve as an even stronger argument
to ramp up the functional fidelity of exercises. These need to be opportunistic and by gathering this significant amount of resources to all serve as training aids for the incident commander, the massive scale of coordination necessary for simulating all those individual parts becomes unwieldy.

This observation highlighted a number of logistics problems with large scale joint training exercises in emergency response. As related by W.C.E., an interviewee in Chapter 3, there is a military training adage goes, “You don’t have guys running on the ground when you’re trying to train the guys in the tents.” The military overcomes this through the use of command post exercises and computer assisted exercises (CPXs and CAXs respectively). Interestingly enough, when talking with fire instructors FEMA/DHS often play down the use of CPXs and again, without following their guidelines, funding becomes problematic. Fire trainer descriptions of the functionals
they conducted varied drastically from similar military functional approaches. The military approach is discussed further in Chapter 3, but the lesson learned from Kings Mall is that there needs to be explicit separation between what constitutes field training exercises from more functionally grounded command post exercises and if one were to organize both, they must be explicitly planned for given the significantly different functions to address. In this sense, the stakeholder goals were lost between planning, the tabletop, and execution. What needs to be kept in perspective, given
all the politics, logistics and costs associated with pulling off such a big exercise— one is never going to have a perfect learning lab exercise, it is not achievable. But this poses a perfect resilience point— support must be provided to be able to consider as many dimensions of the learning lab as one can and more importantly, understand the implications of what it means to not be considering certain other dimensions. Kings Mall organizers cut down the number of testing elements by scaling down the challenges for the ICs. As the learning lab approach predicts, when not looking at the challenges functionally, from an envisioning angle this reduction in scale from one group runs the risk of under-specifying the situated challenges for other participants. As the results of this observation indicate, which are consistent with a previous bootstrapping observation, we saw OCs resort to a significant amount of re-calibration in situ to try to re-balance and inject momentum to keep the exercise moving at the unit level.

2.4.5 Areas to Sustain and Improve

From a learning laboratory perspective, planning, staging, and observation in Kings Mall mapped to a number of learning lab dimensions. The event was an envisioned world and incorporated a large number of stakeholders to delineate the exercise. As events and training challenges were developed further, an interim tabletop exercise functioned as an initial dry-run and ‘before-action-review’. However, scaling impacted the exercise’s transition from virtual table-top to physical execution. While OCs showed well-skilled dynamic efforts to re-organize and stage challenges on the fly, the cognitive training elements and their test:target mappings became arbitrary and often not grounded in original goals and purposes of the exercise.
Over a period of about 6 months following the exercise, we were able to revisit a number of these themes and work directly with the Center. The goal was to take learning from the CTA and envision how we could train up the fire scenario designers to take advantage of Army expertise in CPX design coupled with CSE principles for scaled world studies. Our experiences from the full scale exercise directly applied to how we planned this redesign. We had the audacious goal of wanting to create trainers grounded in CSE principles that emergency response command across the country would want to use. The idea was to use a recently opened EOC to conduct CPXs for urban fire incident commanders. We identified the lack of C2 training support at the operational level, as witnessed at Kings Mall, and developed initial outlines for how we would conduct a proper command post exercise (CPX) with CSE stressors to introduce unintended consequences and unanticipated breakdowns in a more theoretical manner, not just opportunistically and ad-hoc like the instructor injects.

This began with collaboratively establishing an envisioned world from which the CSE researchers and trainers were going to involve students from one of the Center’s training classrooms to experience. This would then lend itself to classical experimental manipulations: does a treatment group that experiences the CSE exercise operate better together? When transitioning to a new environment, do they detect anomalous events differently? Unfortunately, given research commitments and lack of financial resources, we were not able to pursue this line of work to its fruition. However, the process did bring out a number of compelling threads and considerations on how we as CSE engineers would envision running a large scale exercise to facilitate learning. While this study did not materialize due to these conflicts, there was still a high generalizable learning value of this experience which resulted in new patterns of work.
being integrated into the CTA/CWA. Further more fire trainers were exposed to and saw the value in a number of CSE techniques. As well, new design seeds and usability documentation was prepared for the department to help them independently evaluate future software from a decision centered criteria (Trent et al., 2008). Although these dimensions were never able to converge in a single large scale exercise, on a larger level, it was from reflecting over these planning attempts that made us ask ourselves how we would run such a study better. This reflection shaped a number of dimensions of the learning lab approach.
Table 2.3: Kings Mall as learning lab element overview. Instances of observation are marked underneath the five learning lab concepts.
2.5 Recap of Observed Learning Lab Themes

In reflecting on these two exercises, a number of common support functions for exercise planning and learning begin to emerge. In Huachuca, there were a number of interesting scaling and learning dimensions that supported and challenged learning on multiple levels. The observation has this proto-scaling up quality in the sense that it was concerned with observing simultaneous teams. Even though the teams were working the same case in parallel, it served as a dry-run from a research point of view for studying non-co-located teams. It acted as a warm up for doing such an observation with four separate teams in separate locations that have some interdependencies. Certain components played critical functions throughout the scenario that kept the run-time working. The ‘brief the commander’ meetings were fixed periodic landmarks that proved to be quite functionally important in the multi-day exercise. Given that these large scale exercises are longer than the typical one hour/two hour cases, these briefings gave a rhythm and were key functions in structuring the exercise over time. These periodic landmarks established a key rhythm to pace the multi-day exercise. These touch upon a recurring functional theme across all the exercises, which is the need for control of tempo and pacing. Because we cannot simply stop time (as with a simulator), the design and use of pacers and landmarks to move the tempo is critical.

Different connotations of learning and technology played out in Huachuca as well. Huachuca did not have a new technology angle, however, it was designed around learning and to expose students to new analytical techniques. The pooled set of instructors served as stakeholders with the goal of achieving adequate training from the students. The training staff had vested responsibility and given the educational training environment, there is a nominal interdependency that foreshadows the multiple
layers of learning we are interested in. This was related to teaching and it gets back
to the point of contact being a local pragmatic training purpose- the larger long term
organizational learning was related to issues of teaching and understanding better
analytic process (Heuer 1999). A long term learning goal here, in general, is looking at
how to teach people the right way to approach teaching analytic process, independent
of training these specific MI analyst professionals. Usually we say this local training
and longer term training, the local and sharable, is an evaluation of a new formal
process and we tend to think of them as identical. They are not really- if one wants to
learn more about what constitutes good analytic processes one wants to be teaching
people. In this same vein, the Kings Mall urban fire exercise was explicitly focused on
leveraging learning opportunities tuned to the future. Interestingly enough, Huachuca
shares a similar abstract learning goal to this in that the educational exercise was
supposed to function for the (near) future threat environment analysts and firefighters
would soon be working in. For the firefighters, our initial leverage was based on
evaluating a specific piece of new technology (an electronic command board) as well as
new training techniques and procedures. For the analysts, the technology in this case
was in the form of multiple analytic approaches. The intel community is highly vested
in researching, reflecting upon, and sharing what constitutes good analytic process
as well as and how to teach it. The learning goal is sharing that knowledge and not
simply to train up analytic professionals in analysis (i.e. ACH or sense-making). From
a scenario design perspective, we have identified what served as some of the key probes
and pacers that played out, however there was a fundamental limitation in each design
between how stakeholders planned the events. In Huachuca, the stakeholder planning
goal was to cater the exercise as a means to teach and evaluate use of a new analytical
technique, ACH. While it was very interesting from an analytical perspective, the number of challenges we found embedded into the process itself, could be deemed to have constrained learning. From a CSE practitioner perspective, a more interesting goal would be to determine instead whether doing ACH is a good technique. One would want a scenario where doing ACH might be counterproductive, or in which it might slow analysts down, or it might force them to have two hypotheses for too long that are completely unfeasible. While this observation enabled us to gain insight and understanding into analytic process, learning from it also left us better poised to now think about ways to instantiate such challenges into more robust scenarios.

Kings Mall reflected the changing world of urban firefighting and the hypothesized envisioned threats within it. The study was future oriented and that establishes a larger learning context. The future was broader and given new threats and new information technology, the stakeholders faced fundamental questions regarding how to accommodate and take advantage of change opportunities. These are the advantageous futures that are prime leverage points for CSE- it is not just the envisioned world, but rather it is the learning purposes that represent future orientations that go beyond simply just preparing people to work professionally in the world. However, what did we see at the run-time in Kings Mall? OCs had to resort to throwing in injects to make the exercise more interesting and probe groups on the fly. At the run-time, the injects served as probes. Functionally this highlights a key difference between probes and injects. In explaining the OC inject actions, these were injected during the run-time to keep participant interest and engagement. They functioned as probes but they were injects with multiple roles. They were done ad-hoc at run-time by experts because they observed the scenario was too slow, too shallow, and too un-involved.
However, they also served partly as pacers to keep people engaged. While observation is often about exploiting opportunistic moments, from a learning lab perspective the ad-hoc probes were more of a degradation from planning intent not tied to specific goals and objectives. Probes in general were more planful in Huachuca, while more ad-hoc in Kings Mall. While they still served as valuable landmarks in situ to evaluate cognitive work around, from a planning perspective their use and how they were included provides additional insight to planning processes. That being said, both exercises shared many elements of successful learning laboratories which in spite of these differences, contributed to greater understanding about the underlying cognitive work in the domains. The last exercise to be discussed, however, demonstrates a lack of resilience across many of the same dimensions that made Huachuca and Kings Mall successful.

2.6 Strong Angel III

2.6.1 Overview- A Really Big Show

The two exercises described above contain a number of elements deemed important for learning laboratory exercises. In both Huachuca and Kings Mall, CSE researchers tended to be involved early, training elements (intentionally and not) exposed cognitive work, and the exercises were of a sufficiently manageable scale. In contrast, Strong Angel 3 lacked a number of these elements and thus, not surprisingly, despite vast resources it was a much less effective exercise than the two described above. Strong Angel 3 serves as a cautionary tale and was the motivating experience for us to begin to develop a rationale for why large scale exercises should be designed for learning. While the exercise meets a number of scaling dimensions, given the significant absence
of key learning laboratory components, very little useful learning, let alone data, was obtained.

In August of 2006, CSEL was asked to participate with the University of South Florida (USF) Institute for Safety Security Rescue Technology in the third Strong Angel integrated disaster response demonstration. This event was held over five days at the San Diego Fire Training Academy in San Diego, CA and incorporated over 800 participants from 200 organizations including first responders, NGOs, military officials, exercise observers, and commercial, university, and military researchers.

The purpose of the demonstration was the development of a laboratory for experimenting with cutting-edge techniques and technologies to facilitate improved cooperation and information flow across the civil military boundary in post-disaster and post conflict environments. (SAIII final report, 2006)

Strong Angel 3 (SA3) was a major step up in scale from the previous two observations discussed in this chapter. SA3 was a break down in emergency response exercise planning and we can take this problematic performance and assess how it broke down across multiple levels. SA3 serves as an exemplar event from which to learn from in order to better define and develop large scale exercises. We could analyze SA3 in detail, but this runs headlong into the problem of being seen as a negative attack. Looking at it from the perspective as a real coordination event- the goal here is not to envision how to solve an emergency scenario and contain an outbreak, but rather the goal in looking at SA3 is to learn how to design the next crisis response learning experience better.

The overall goal of SA3, ostensibly, was to test civil and military social and technical systems in a simulated disaster response. This was heavily geared toward mutual developer and practitioner learning about current and near future communications
technologies. Right off the bat, one should be concerned that this might be way too narrow a scope for such a huge exercise. As any practitioner or researcher intimate with the field would relate, communication is only one small piece of coordination in crisis response. Bringing 800 people together and planning this huge event simply to share communication activity was a crucial underestimation and a mistake for planning. This key consideration serves as the first canonical example of similar latent under-specified decisions made during planning. After the skeleton of the exercise was devised, the next stage in planning for the exercise was to bring in a number of stakeholders and outside researchers to propose respective goals and task areas of interest. From an envisioned world perspective, this would appear to be an excellent way to avoid parochialism, however as will be seen later, the way outside perspectives were incorporated, instead amplified this problem. An executive planning team was responsible for coordinating the goals and purposes of these various researchers and then develop opportunities within a realistic disaster scenario that would escalate in similar ways to an actual disaster. When crafting this scenario, the planning team left open opportunities to include new challenges in situ for the various participant groups to face. Strong Angel was designed to be a showcase for researchers, system designers, and the emergency response community to work together and learn about new communication technology support for future response conditions. As can be seen in Figure 9 below, all did not go according to plan.

Strong Angel 3 was a major shift in scale and scope for such emergency response exercises. Based on our own scaling dimensions- the exercise implementation adheres to each level of scale. However, in contrast to this degree of scaling their initial learning goals were very short sighted and did not expand across these same dimensions. First,
the exercise spanned across a large physical scale. Sorties traveled to multiple locations across the city and to the outskirts. In terms of work scale, multiple groups both co-located and distributed had to coordinate across the cold, warm, and hot zones that made up the large disaster area. With regard to the temporal dimension, the exercise played out for 9 hours a day, for five days. Overall the event involved a large number of stakeholders interested in the future of disaster response and management. The explicit backdrop for the event was based on concern in the community that there needs to be a future given the changing threat environment emergency responders were operating in. It was staged and resulted in significant participation because of
the growing concerns in the community acknowledging their world is changing and
demands are increasing. This acknowledgement of underlying capability change was a
key driver for conducting the event and the purpose from which groups volunteered
their efforts and resources. The community wanted to learn how new technology
could support their work better. However, when it came time to invite researchers
and new stakeholders to establish the more tangible learning goals, planners isolated
each individual project and did not bridge these goals into a larger context within
the exercise as a whole. A white cell was supposed to take some of the burden of
this responsibility, however with no top-down planning and the active exclusion of
bottom-up teams at this critical stakeholder planning phase, the stage was set for
learning breakdown.

There was no CSE input into initial planning of the exercise. After the initial
planning process our CSE team was brought in and asked for research suggestions.
Our team had originally proposed two task areas of interest to investigate. The first
task area was to examine command and control scaling efficiency and observe social
dynamics within urgent environments. A distributed observation plan based on the
Huachuca methodology was formulated, taking into account the order of magnitude
difference in size between Huachuca and SA3. Probes were not known ahead of
time and the exercise committee kept the scenario under wraps. This made it quite
difficult to pre-plan and pre-position observation resources. A secondary task area
involved coordinating multi-sensor information streams from deployed search and
rescue robots for a cross-country medical reach-back experiment. As a participant
member of the rescue robotics group, we were interested in how access to remote
visualization and reach-back coordination communications changed the fundamental
nature of the response work. The goal was to experiment with different methods for intelligent bandwidth regulation of wireless communications for the tactical response robots. The PIs served as both participants and observers across these two areas. Unfortunately, despite significant pre-planning on our part attempting to prepare to be surprised come run-time, we saw first-hand how a number of disparate forces can lead to disaster in large scale exercise design and execution.

2.6.2 Scenario- A Touch of Real-World Complexity

A highly lethal (simulated) global pandemic kicked off Strong Angel 3. With no vaccine available, quarantined areas were established and began to pressure emergency and disaster response centers. It was against this backdrop that a terrorist cell launched a series of cyber-attacks on a medium sized city. Electrical power, phone service, and internet connectivity were all compromised. This served as the narrative that was alluded to when the exercise started. Like all good CSE scenario design techniques, it started in medias res. This initial confusion was intended to pace and organize the multiple teams that had gathered for the exercise, however the narrative was soon abandoned as difficulties began to escalate from the initial challenges. Our team was situated in one particular response center set up in a cold zone just outside a quarantined portion of the city. The entire center was made up of a diverse group of technology developers, vendors, first responders, and researchers set up to demonstrate, field, and learn how to effectively exploit communications and respond to cascading disaster situations. Across from the center was a command team of National Search and Rescue personnel (NIUSR) responsible for coordinating field exercises and pushing the scenario forward with the exercise committee- a hybrid white cell and response group.
Their command operation was set up and conducted from within a secluded closed building in a hand-shake free zone. Little information went in or came out. A third area was set up at the entrance to the site, where a decontamination tent was situated behind a sparse contingent of border patrol and military units. These units were often gone with no status of return. In between these three islands, multiple trucks and vans carrying some of the most sophisticated radio, satellite, and mesh networking communications hardware available were parked and powered without interruption. Two large dishes both pointed to the same satellite- theoretically providing a rich 10 mbit WAN for the entire response operation. Technicians feverishly worked for four days to bring all communications online. They never succeeded with this task, which came to serve as a massive stumbling block to get the scenario running. The only solution the planning team and communications groups repeatedly instituted was to shut down all systems and forbid any computer usage (networked or not) on the site.

“CISCO has enough to take the load.” said on Day one.
“Were shutting it all down.” said later on Day one.
“Shut off all encryption.”
“There are too many applications, protocols, and devices running we know nothing about.”
“This is unprecedented, the users are usually the last to know theres a problem!”

Over the course of five days, network connectivity never recovered, field sorties were drastically limited and became increasingly ad-hoc, and the principle concern of the groups in the cold-zone was not isolating the pandemic or venturing into the warm/hot zones to gather information or perform knowledge elicitation with practitioners, but rather, many simply waited for network connectivity to return so they could pull in timely available data in order to, as they perceived, help everyone ‘coordinate better’. 86
Our own observation plan was designed around pre-positioning multiple observers within the planning committee and the NIUSR white cell and then moving additional observers to opportune information rich locations. This became impossible given how the exercise broke down which goes far beyond just the accident phenotype of a communications failure.

2.6.3 Scenario Execution- How it broke down

Almost a year prior to the exercise run-time, the executive planning team was faced with a daunting challenge for the SA3 demonstration. They spent many months planning for a 100 attendee event and ended up being inundated with over 800 participants. Given this size increase a number of scaling issues exacerbated how the demonstration was compromised. Strong Angel III came to resemble a microcosm of a disaster response itself. For example, the cyber attack scenario element was originally introduced as an excuse to stage a network infrastructure build up. This is a common scale-up problem in real disasters. Much like in the real world, however, network connectivity and communications were the first thing that failed... and never recovered. The executive team quickly acknowledged that the original plan needed to be modified, however, the team had no ability to make these changes because of the weight of the exercise already in motion. The team had planned for the role of an emergent synergy operation, named Shadowlite, that would function as a white cell for running this. This team, formed by members of NIUSR, was supposed to be responsible for dynamically generating the data and scenario events and injects to steer the dynamic course of event at run-time. Instead, Shadowlite functioned as an independent command and control cell for conducting their own field exercises opportunistically and independent
from planned. They did not disseminate information to keep the exercise moving or share data with other groups and ultimately did not allow exercise observers to participate with them. This degree of information hoarding is exactly the opposite of what a white cell should be responsible for and further undermined the critical sharing function they were meant to fulfill. What should have functioned as a key observation command center was literally closing its own eyes and ears by refusing to work with observation teams. On top of this, general leadership was not available to either integrate this team or support the validity of the assigned observer roles. In the absence of such leadership providing any cohesion, the group pursued its own agenda.

The communications breakdown significantly fractured participant groups from the start. While some participants, including those in our own group, believed the uncertainties were part of the exercise, it quickly became evident that they in fact were not. Planned synchronized group activities were abandoned, while the small leadership team told research and vendor groups to hold off all exercises until network communications were fixed. On Day 2, shortly after the outage, the exercise leader was overheard saying that comms would be back up in 20 minutes. Comms never came back up for the rest of the exercise. By continually delaying and deferring events, it became quite challenging for both participant and observer groups to develop an overall coherent view of events. No planning team or independent group was able to assemble this big picture roadmap. In this multi-day period of down-time, motivated sub-groups formed to stage their own exercises. This form of adaptation did let some groups exercise their equipment, but these situations were often formed on the fly and without notification to observer teams, the white cell, or leadership. This led to many groups getting basic usability and some usefulness feedback due to the grittiness of
real world operations, but this local learning (at best) had no avenue or outlet to share. There were minimal opportunities for reflection across groups to share and discuss what was learned. Organizers originally wanted to throw challenges and structured opportunities at sorties to gauge both new technologies and incident management systems, however, this suffered greatly given these breakdowns. By the end of the exercise on the last day, for the most part sharp-end first responders were just as distanced from the technologists as they were at the start. Despite all the pre-planning, we observed multiple competing and conflicting goals when it came to execution across leadership, stakeholders, and participants. On top of that, the lack of leadership did not reinforce planning goals or maintain a coherent exercise framework. With no contingencies in place at execution, whatever local learning from collaboration across disparate groups was gained could not be shared or generalized given there were no opportunities for reflection or synthesis for this kind of learning.

2.6.4 Lessons Learned

Strong Angel III was a future-case field observation and participation event that illustrated the challenges of conducting large scale exercises while balancing research tradeoffs in a technology oriented exercise. As related in our own after action review of the event, this large scale exercise did not simply fail because of a poor scenario or brittle execution, but from coordination breakdowns across multiple levels of stakeholders and participants resulting in fragmentation. At Strong Angel, no explicit leadership organization group was present and operating at the exercise. On paper there was a scenario, a relatively face valid scenario of flu pandemic and cyber attacks, however when it came to run-time, there was no explicit plan, no determinable organizers to
talk to, and no real pacing that updated the state of the simulated world. A large amount of this came from the massive (unplanned and irrecoverable) communications failure early in the exercise. \cite{Woods2006} emphasize that the critical component of any staged world study such as this depends on designing the scenarios participants face- Strong Angel set the scene for just that and then merely stopped the show as soon as the proverbial curtain went up. When their exercise plans came in contact with the real world, they did not survive.

Organizers began to design problems, however these were never expressed in the scenario or shared ahead of time. There was minimal continuity and momentum of the scenario, and as soon as it started to fall apart there was no pacing to move things forward or ways to recover lost ground. With the absence of pacers and contingencies, coordination broke down and people began to splinter and pursued their own agendas. Such adaptation should be the starting point of an exercise around which cascades and escalations can be built-in, not something to be deterred and chastised. This also highlighted gross lack of consideration for the roles of observers. As mentioned, there was no meta group coordinating observation, and all observers were representatives as PIs of submitted research plans. On one level, the lack of any deferred authority to observers to coordinate with planners or the white cell (who performed no observation throughout the exercise) to adjust the scenario could have been useful. A much bigger observation issue arises, however, when the entire observation infrastructure was independent and self selected to research their own interests, it starts to become very difficult to impossible to expect observers to discover new patterns of work when the blinders are built in- almost a Psychologist’s Fallacy effect \cite{James1890,Woods2006}. Not only could observers not coordinate, but they were
already looking for pre-determined criteria without knowledge of the actual exercise, and furthermore there were no opportunities for reflection with other groups or with different perspectives to broaden in these activities. This is the exact opposite of the critical lessons learned from Huachuca and Kings Mall regarding the critical role observational agility played.

As mentioned earlier, a number of competing and conflicting goals within the planning community were exacerbated by the scale and scope of the event itself. These goals were poorly articulated and a number of different expectations and assumptions based on technology were either brittle or never materialized. Stakeholder roles and stances, and their goals and positions, respectively, were markedly mismatched throughout all stages of the exercise. The events and demonstrations that were attempted were often ad-hoc. In discussing the Event Based Approach to Training (EBAT), Fowlkes, Dwyer, Oser, and Salas (1998) cite the profound importance of taking a principled approach to incorporate training objectives, exercise design, and performance assessment in such real-world simulation training. In EBAT for example, scenarios are designed around introducing events within training that provide known opportunities to observe specific participant behaviors. Control of the nature and number of events helps preserve this explicit linkage. In transitioning EBAT for military training of large distributed teams, Fowlkes et al. (1998) saw that scenario control rapidly becomes problematic. Much like in Strong Angel, these cases involved hundreds of participants and teams each pursuing their own goals and objectives. Fowlkes et al. (1998) contends that the same elements that make such large scale exercises so desirable, environments that are complex, dynamic, and unpredictable, are the same factors that make them so difficult to create and train for. In attempting
to overcome this in one particular application, the researchers found that trainee interaction led to many unplanned events. They cite this as the need to increase observation so instructors can keep events on track, however from a CSE stand-point, such interactions the learning lab would consider as valuable sources for observation of adaptations. In either case the same message is clear, increases in complexity require necessary increases in observation control.

As a future technology event, the initial SA3 planning mindset fell into the trap of constructing the exercise in an overly technology-centric point of view, the very problem envisioning cautions about. Similarly, we were invited to participate based on our work with and access to rescue robotics, not as experts in human evaluation and anomaly response. Rather than pose exercise challenges based on the underlying challenges of what makes response and management complex, they focused instead on how new technology should work. This is not novel and is a typical illustration of technologist/practitioner gaps (Woods & Hollnagel, 2006). In order to envision potentially useful applications of technology, system designers must have a thorough understanding of the underlying cognitive challenges faced by practitioners and how those challenges would be supported or hindered by the introduction of new technology systems. To oversimplify, the practitioner is often focused on solving his/her immediate real problems independent from the state of technology. The technologist, in contrast, is often focused solely on developing new products, nonaligned with actual practitioner needs or input. Thus a gap is created where the technologist is left unable to understand how his/her systems can help and the practitioner is left in the dark with regard to what the technology actually does or how it could help. This dichotomy often exists at stakeholder and participant levels and tends to result in exercises designed only to
test specific technologies or exercises that only test practitioner core competencies, and rarely in the middle do they both meet. When this gap still exists at the end of an exercise, as with SA3, no learning has taken place. We can learn from Strong Angel’s shortcomings and coupled with the emergent themes across all three of these observations, synthesize better approaches to large scale training.

SA3 was impressive in the sense that the planners overcame significant logistics challenges and assembled over $35,000,000 in the latest communications technologies to test advanced forms of communication in disaster response. However, given these planning shortcomings it should come as no surprise it fell apart. CSCW as a field in general has shown that simply having communications infrastructure is not the same thing as having effective synchronization and coordination, especially when cascades of events are happening over multiple temporal and physical scales (again, the escalation principle). These cascades are stressors to resilience. Many technologists and practitioners after the event still felt that the fundamental problem was the need for more reliable communications, and that represents the greatest learning failure of all. A resilience perspective, from the start, would posit that you will always have communications problems and breakdowns and it is being able to support the adaptation around such problems that is critical. As with Three Mile Island- even with perfect communications connectivity and everything working, coordination did not come for free- operators had no idea what to say and to whom as events escalated given the significant understanding level gaps and disparate data sources. SA3 resulted in a number of dead ends in this sense. Most of the pre-planning centered on large scale topics- i.e. power and communications and establishing connectivity. Companies brought large scale technologies- IT to be used at the leadership level and two dedicated
satellite streams. With all this emphasis on communications, as soon as a little real world complexity crept in, the true foundational problems with the event tore open. Organizers and stakeholders need to thoroughly plan and envision how to interject technology in place with different aspects of the scenario—this is exactly why CSE puts such effort into deliberate planning in envisioning exercises at the front of the exercise while also planning to anticipate changes throughout.

A common critique from multiple observers at the end of SA3 was that it quickly became difficult to discover what would be useful for training or adopting new technology. We can see that this was not just due to a single breakdown in the exercise, but short-sighted and underspecified planning throughout each stage. To contrast, in both the MI Huachuca case and the FD Mall exercise skilled and veteran instructors not only displayed ready ability to modify behavior and adapt scenario elements based on their perceived changing circumstances (some more planful than others), but such participant adaptations were accepted and in some cases encouraged. Functions existed and were built into these exercises that avoided such breakdowns as seen in SA3. Observer controllers had deferred authority to steer events while remaining tightly coordinated with the planners. In SA3, on the other hand, the staging and attempts to recover from the initial breakdown left no way to either integrate other technologies into the flow of activity even at the micro levels of individual groups or enable other groups to deviate and conduct their own activities, even after recovery was a lost cause. The planned scope of observer roles was relegated to a very limited perspective and with no integration to overall planning or deferred authority, significant learning opportunities were lost. How could a learning lab approach have prevented this breakdown? There is nothing inherent in the learning lab that would suggest SA3
could have been a resounding success. However, contingency planning is a critical element that the learning lab stresses. This is also consistent with the theme of a resilient observation system. The massive communication breakdown should have served as an opportunity for discovery. The proper approach should not have been ‘stop doing everything’ but rather to adapt and adjust the scenario as if, for example, the cyber-attack had spread and compromised the satellite communications. A resilient system would have realized that this is the exact kind of thing that would happen in the type of disaster scenario being studied. The first thing that would happen is the comms break down, and at SA3 we essentially got that one for free. The problem was, however, given the parochialism and short-sighted envisioning, the planners had envisioned this as an exercise to showcase specific vendor and stakeholder technologies. Such a change however, was a totally unacceptable scenario not just because of the lack of planning authority, but because of fundamental goal conflicts with stakeholders that should have been addressed at planning. Whereas brittle technology could have provided an opportunity for discovery, fundamental conflicts throughout planning and execution prevented such success.
Table 2.4: Strong Angel 3 as learning lab element overview. Instances of observation are marked underneath the five learning lab concepts.
2.7 Synthesis Across Field Observations

In the world of disaster response and preparedness, organizations are running exercises similar to these three studies more and more often. Given the changing threat environment, organizations and participants want to learn from such events, but as we have shown with the small sample of exercises here, cultivating learning opportunities from large scale exercises demands coping with significant expenses, logistics, and scaling costs. As exercises grow in scope, the scaling logistics make it difficult to run these exercises multiple times. These large scale events have one shot to deliver and are further complicated by the introduction of multiple groups, events scaling over time, and challenges scaling over physical space. Based on the descriptive meta-analysis of these exercises, the goal is to aggregate the key exercise components identified above in these observations to extend the learning laboratory concept into a working framework for the guidance of conducting large scale resilient exercises. Given that participants and stakeholders are not getting all the value out that they can despite the significant staging costs, the learning laboratory framework lays out the tangible elements necessary to maximize value and learning yield.

The preceding analyses shed light on why components of these exercises fell short of CSE goals and we can begin to extend the approach from concept to framework. The first two exercises are somewhat consistent with a learning lab approach, but the last exercise, Strong Angel, violates both the broad perspective of a learning lab approach as well as multiple detailed local elements. The learning lab framework extends the CSE approaches for doing large scale exercises by taking a functional approach to incorporate components of what worked from what was observed in the exercises as well as building in resilient functions to cope with uncertainty and exercise brittleness.
This synthesis serves as a way of incorporating these lessons learned as well as provides an understanding of their shortcomings and failures. Synthesizing across the summary tables and reflecting on the lessons learned from these observations, a few general patterns of cognitive work emerge along with a number of scaling and learning issues. In Huachuca, a variety of cognitive complexities were seen that made it difficult to carry out analytical work. These same complexities have since been used and re-staged to illustrate these same challenges in collaborative workshops. From the urban fire exercise, long term learning efforts have used these results to aid in developing a large function based cognitive task analyses for urban emergency response organizations. Last, in Strong Angel, the profound result was that the disaster event itself was a multidimensional breakdown in learning and serves as a rich lesson for exercise design in general. In reviewing these observations, it becomes clear that a number of learning opportunities in other dimensions were under-yielded. Given the multiple scales and dimensions one can invest in, one needs to be able to assess how to maximize learning yield given the scale of investment and various costs that go into pulling these off.

Future and envisioned worlds present new challenges to what constitutes expertise as well as how groups coordinate in the face of new kinds of threats that could complicate incidents in a response. New threats raise questions about what is expertise, what is coordination, and what is resilience. As we move back toward staging scaled world studies, these methods tell us to set up conditions to look at a certain places and then let the world tell you what is interesting. Observations are then reported by organizing around the concepts of interest along with why observers thought it was unique to look there. In working the envisioned world problem, the results of any work domain or task analysis should strive to explain observations grounded in the cognitive and
cooperative activities seen in the current field of practice. This will provide insight toward how to inform or apply to the design process since the introduction of new technology will transform the nature of practice. Investments are needed to support those new roles and demands, and to mitigate new forms of error. We gain leverage by shaping the conditions of observation. In the three exercises, pointers were made regarding what we might change to address shortcomings in execution or challenges to cognitive work. But how do we design such events from scratch? We can create ecological models of the process, embark on scenario design, and discover and harness artifacts as probes and disturbances. Functionally though, how can we build this? The focus can be simply local learning such as evaluating the usability of a specific device in a specific context, all the way to generalizable learning—creating understanding with regard to the larger dynamics that play out in these settings as natural laboratories of cognitive work.

When organizers do not follow a solid plan for doing this, as with SA3, there are missed learning opportunities. Organizers need to re-evaluate traditional approaches such as those outlined in Chapter 1, because the world has moved on from these relatively compact exercises to complex co-located teams dealing with specific design and development and human/team performance issues to distributed systems issues, multiple locations, synchronous as well as asynchronous interaction, and to large scale team performance. Strong Angel’s breakdowns in particular read quite similar to the breakdowns seen in classic expert system design in that: plans for execution were under-specified; there were clearly bugs in the plan; novel situations were not planned for; and ultimately, there was no contingency to deal with the multiple failures (Roth, Bennett, & Woods, 1988). These are the types of challenges that must be mitigated
for successful staging and learning. Across all of these exercises, there tend to be classes of functions and events commonly built into them. We see specific elements that serve as probes and we can identify how temporal rhythm is influenced by various pacers and scenario landmarks.

This type of deconstruction is critical especially when one does not have complete control over the scenario design. These scenario and event maps must tell the observers what they need to be prepared for because these various classes of elements and events are supposed to happen. A specifically challenging illustration of this was faced by [Watts, Woods, & Patterson, 1996] when studying mission control. Watts was confronted with a week long anomaly and had to deal with a number of general human factors issues that needed to be overcome. The most notable was the problem of needing to access expertise on demand. In Huachuca we built up familiarity by experiencing the exercise itself and exploiting the ability to bring in expert commentary as needed so as to continuously update a projected map of the exercise. With the NASA case, it was large scale and it was an actual event, not an exercise. The observation took place over multiple days and multiple teams were involved in the anomaly. To deal with the distributed multiple teams, given only one observer, teams could not be forced to talk-aloud or be interrupted, so effort was dedicated to studying one team in particular and understanding their perspective from which to broaden and observe cross team interactions. By richly tracking one team and observing how teams got together for coordination meetings, an understanding was built based on how others worked. This particular case is interesting because of the uncertainty built in to real world events, there is no master scenario or ability to know where things will go next. Being able to capture what was interesting in spite of not having the ability
to have landmarks and probes, forced the researcher to have to deal with events as they came up and notice what is interesting. This is a general issue in any found case observation where in the absence of building up domain expertise, one must be able to pinpoint deviations and bring in experts to bear quickly. Whether real or simulated, there are significant challenges to conduct good observations. These challenges demand flexibility and the need to have coordinated observational agility. This mission control study performed by Watts et al. (1996) in particular illustrates problems of trying to track multiple organizations or sub-organizations within a large one, and how to you coordinate a research team to understand those interactions. These functions come to the forefront as we lay out the specific elements in the learning laboratory framework below. The functions, however, are only the interim elements of the exercise bookended by input from the goals and purposes established by stakeholders, which result in output as various degrees of learning. Before we can build such a framework, we need to understand why these mappings are necessary and how such mappings play out in real world studies.

2.7.1 Mapping From Test to Target

To get generality from field observers, researchers need to make the test:target mapping explicit (Woods & Hollnagel, 2006). Being able to identify what observation or probe samples what aspect of work or represents what instance of an activity, is often a many:many mapping between physical and functional components of work and represents the fundamental reason we cannot and should not conduct exercises like the experimental psychologist alluded to earlier. The level of shaping of the conditions of observation can differ immensely, however a fundamental rule for sampling the
world in ethnographic observation depends on accurately describing who we looked at, what probes were injected, and what conditions were set up that would generate probes. Situated performance can then be organized around probe events relative to a standard paths of performance (i.e. canonical paths). Observers, and planners, need to be able to quickly recognize where something different from planning is going on or to be able to quickly identify where something might need more elaboration or investigation. Especially in found cases, this is what prepares observers to be surprised and identify probes in situ. In Huachuca for example, observers assumed certain kinds of communications were going to work, but then when they broke, we saw these as probes. Even though planners and organizers thought they were testing a multitude of collaboration tools under the assumption that these new communication modalities were going to be highly reliable, the fact that they broke down turned into a different kind of test than they thought.

From such an approach to observations, there are a variety of general ways to do protocol analysis and process tracing to abstract analysis further. We can expand upon this given large scale exercises. Because of the scaling complexities, events and adaptations often happen differently from what people expected. As we saw with Huachuca, running the same event with multiple teams produced different results in each run. Varying small details accumulate in themselves and in ways people respond to them, to create a qualitative shift in cases. The case isn’t the same anymore, one has to think of it as an envelope of paths. It is in the rich contextual observation that insight is generated by being able to create, track, and analyze these envelopes. Across the three exercises above, similar techniques and methods behind the observation and analyses were employed extending the basic process tracing approach.
The process trace, as a general technique often used in CSE, is what gives us the most leverage and helps researchers identify these envelopes of fluency and adaptation. Generically, based on observation data, researchers start by taking a general functional stance and then performing a process trace to run through and tease out cognitive functions. Sanderson’s design efforts such as MacShapa and Guerlain’s RATE tool were specifically built to support observers to organize material for such process tracing (Sanderson et al. 1994; Guerlain, Shin, Guo, & Calland 2002). The process trace is not necessarily analysis in itself, but rather an organizational technique to facilitate exploration as a means of organizing massive amounts of raw data before selecting instances and events for further analysis. From a process trace researchers could perform protocol analyses, discourse analysis, or a variety of analytic techniques. For the past 20 years, it is common for observers to have ready amounts of data in the form of video, observer notes, photos, and audio and one needs to be able to coordinate the multiple data streams effectively. Simply aggregating all this data together does not automatically lead to discovery. Abstract process tracing capitalizes on using multiple methods from which we can get more concrete or abstract as to the particulars of the domain. From these dynamic processes, we can relate observations of how mindset evolves to more abstract cognitive functions. At the lowest level we can identify affordances and constraints and look for indirect indicators on this. It is this general organizational perspective that shapes the learning lab so that it starts to take on the characteristic of an exploratory map enabling individuals to look at the cognitive work within a functionally structured set of goals in an exercise which can then lead to synthetic interpretations of the work in context.
We have often said that what drives good observation is the ability to prepare to be surprised. As we approach the learning lab framework, we want to expand this preparation from not only preparing for observation, but preparing to design large scale exercises themselves. Surprises challenge the way of doing things, which leads one to naturally ask, what are the purposes behind those surprises? These purposes create the *why* behind the exercise. When a surprising event occurs, that immediately should trigger the observer to ask, what’s the diagnosis? The suggestion is that certain directions should have specific actions associated with them. Those directions now become new *whys*. This establishes a fairly generic outline- that a surprise triggers a diagnosis, diagnoses propose changes, and this spurs investigations of those changes. In observing and guiding these exercises, one needs the ability to investigate changes given a diagnosis based on surprise, and this provides an anchor for us to create large scale human system performance studies. In human performance studies one expects intricately tangled connections typical of a functional decomposition on a real system. These are not formal ‘neats’ and that is why planning, observation, and analytical support must be built in to support the tractability between goal means decompositions from purposes to learning with this approach. The human performance study now can support more than just the performance evaluation, but rather feed into a multitude of learning directions, toward design evaluations, and toward new innovation- building this learning is ultimately about building understanding. The learning lab as a performance study becomes a design direction study, not just an evaluation, from which to learn and understand cognition at work.
2.8 An Initial Framework for Learning Lab Design

Across the synthesis of these exercises, a number of structural elements and core functions necessary to support learning and studying cognitive work in large scale exercises emerges. As one scales up, costs increase and the logistics make it difficult to stage multiple runs. Large scale staged world efforts require extensive planning and specialized roles in order to balance stakeholder and participant needs along with long term learning goals. An important lesson learned from all the studies above is that if exercise planners wait to start the learning until after they have planned or run the exercise, then they are not going to learn very well. Cognitive Systems Engineers can play a crucial role in this step. From the ground up, the learning lab framework is designed to facilitate and provide insight into learning based on the vital functions identified in the lessons above. The basic geometry of the learning laboratory, as seen in Figure 2.10, is arranged around three phases: the staging and planning of the exercise; the execution and run-time; and the sharing of learning at the conclusion. The relative arrangement of these parts is further deconstructed based on key roles, critical support functions, and the functional mappings between one another.

2.8.1 Key Roles: Stakeholders, Observers, and Practitioners

For these types of exercises, in the Francophone tradition, the stakeholders are the problem holders, and they are also the practitioners (DeKeyser, 1992). As organizers prepare for a study, they must respect the learning of the participants in that field of practice and plan to support this participative mutual learning, knowledge, and feedback from the start. Stakeholders, their purposes and goals, constitute the foundation of a learning lab. Strong Angel served as a prime example of stakeholder
coordination breakdown from the start, however, this is not the only case where organizers were unable to handle logistical burdens of handling multiple stakeholders. As will be discussed further in Chapter 3, we see a similar breakdown in the evolution of stakeholder conflicts in a failed NASA aviation fatigue study. The study never happened because of the difficulties in managing the multiple perspectives and goals. While SA3 planning did not succumb to the same initial logistics and timing costs, it instead did not attempt to evaluate or balance across the plurality of stakeholder and participant goals. There remained a critical need to facilitate and mediate the overlapping and multiple conflicting goals and purposes of the stakeholders. Perspectives from the different stakeholder groups must be externalized and then represented throughout
planning and exercise envisioning. The multiple stakeholder groups have a variety of purposes- some might be technology oriented, some might want to demonstrate a system (again, plurality), some are research oriented. Because there can be multiple learning objectives from these groups, it increases the need for all stakeholders to be involved in initial preparation and all stages of envisioning. These roles serve to establish the specialized personnel and necessary CSE methods and approaches to consider before exercise planning even begins. Throughout this it is critical for stakeholders to have artifacts and forums supporting their ability to share, elaborate, critique, and collaborate to mitigate the four negative dimensions of envisioned world components.

Leadership, as a function, must derive from the stakeholder and problem holder groups responsible for the planning of the exercise. This creates the need for specific support roles to balance stakeholder and practitioner interests in order to set-up and conduct these exercises as learning laboratories. Building off the key support teams in the building block overview in Chapter 1, the first support role consists of an in situ group that runs the exercise and observation on site and caters to the local actors within the context of the exercise- these would be akin to the embedded observers in each of the exercises above. The in situ group actively participates in the event while also making sure local operations run smoothly. The second support role proposed in the framework is for a meta group that is responsible for fulfilling the research base and learning goals. Ideally consisting of cognitive engineers, the meta group oversees the learning objectives of the scenario and supports the in situ group leading onsite operations. Similar to the functional role of the command center coupled with the instructors in Huachuca, as well as the exercise planning/observer controller group in
Kings Mall, these teams provide a valuable top-down cohesive structure. The meta group must monitor and coordinate the larger observation and learning objective efforts which frees up the in situ group to be able to focus on observation. The meta group functions to monitor all observation and decision processes, including their own, in order to detect when the scenario elements and exercise may begin to drift toward boundaries. The meta group is ultimately responsible for tracking and conducting exercise observation employing (multiple) CTA techniques (i.e. functional analysis, concept mapping, critical incident, future incident) and injecting the next or adapted scenario events based on collaboration with the in situ team and field observers. They
must be detached to ensure research goals are being fulfilled and work with the in situ
group if goals are not being met, the scenario tempo needs adjusting, or events must
be changed.

2.8.2 Staging, Set-Up, and Functions

The staging phase of the learning lab is first broken down into staging set-up, where
stakeholders bring to the table multiple purposes that lead to a variety of functions
that have to be integrated into the exercise- a plurality of views. Given the challenges
of envisioned world planning, care must be taken to balance across the four dimensions
(plurality, underspecification, groundedness, and calibration) and determining what
the practitioner and research goals of the exercise are going to be. These form the
foundation of the learning lab, the Purposes and Functions. Avoiding the goal conflicts
and finding alignment is not trivial, especially as seen in Strong Angel. A specific
technique described by Klein (2004) describes the use of the pre-mortem as a crucial
technique to use in business planning sessions to avoid such planning breakdowns
and conflicts. The functional goal of a pre-mortem is to find weakness in a plan.
The pre-mortem takes the format of group critiquing to identify weaknesses through
mental simulation. The group collaboratively attempts to identify weaknesses and then
assess ways to counter them. Individual team members first familiarize themselves
with the overarching goals, they then imagine worst-case scenarios (i.e. envisioning),
responses are generated to those failures, and finally responses are consolidated by
a facilitator (i.e. a member of the meta group). Planning decision makers can then
use the pre-mortem to improve the overall planning, anticipate problems, and cater
specific functions to them.
An envisioned scenario is developed based on the Purposes and Functions and specific learning lab functions are incorporated. In parallel, a script and sequence of events are further elaborated to drive the scenario in relation to these functions. As the script is further developed, it starts to become a physical artifact that synthesizes and integrates the multiple stakeholder purposes. This is then incorporated with the multiple learning lab functions to be fulfilled in the exercise. These functions scaffold the scenario along with in-situ team operations, and are developed in parallel with the exercise script. The initial four CSE-grounded functions to be defined below are in the form of:

- Probes
- Pacers and landmarks
- Rally-points
- Reflective forums.

In the scenario design process, embedded probes are injected within the normal flow of activity that do not require interruptions to the ongoing scenario. Probes are placed around boundary conditions by having situations escalate to create mismatches between the systems behavior and the agents expectations- for further description of probe events see [Patterson, 2004]. Effective probe design illustrates hypothesized demands for a scenario based around the complexities of the domain. As researchers, we are looking for when and how practitioners can get trapped on the garden path from which new scenario trajectories can be introduced (such as consequences of introducing latent interactions). We introduce probe events in scenarios at system limits and
Figure 2.12: In the staging phase, stakeholders define key purposes and functions for an exercise and then construct an exercise scenario incorporating multiple learning lab functions for the run-time.

boundaries to observe adaptations and later trace the process of change around them. An example of a disturbance class probe event is shutting off a particular device. Such a probe is useful for observing the effectiveness (or brittleness) of a new technology based on the adaptation to the disturbance. In designing scenarios around passive situation probes, researchers can make the complications and challenges to practice observable.

These functions can interact with the scenario to create more dynamic events. Probes and cascades of events can be contingent based on what participants or instructors do- organizers and observers want this forecast ability within an overall
map of the scenario suggesting where events are supposed to be difficult. Maintaining an overall map of where probes are occurring identifies places to focus observation, but given the envelope of responses, a probe might occur naturally that was not identified ahead of time. This is the sort of serendipity that must be accommodated, especially given that increased likelihood that real world variations creep in given the larger scales of events. Organizers and observers must realize that real world variations will creep in beyond the design of the scenario. This proved disastrous in Strong Angel, but in contrast, proved quit fruitful to Roth et al. (1988). In their expert system based elevator study, the researchers did not try to build a novel case, but they got one anyway thanks to the variety of the real world. The researchers were not trying to create a difficult case for the operators, but because they were using a real system they ended up with genuinely difficult cases. In SA3, the downed comms was too brute of a break down, but it illustrated a general problem with their planning (and lack of envisioning of failure conditions). Organizers and planners should have assumed certain kinds of communications were going to work, but then they broke, and that became another probe. Even though SA3 organizers thought they were testing something else with the assumption that these new communication modalities were going to be highly reliable- in fact when they broke down it was a different kind of test than they thought. We would see that as a new kind of probe, however, they saw it as a dead end. In a learning lab, that does not mean that one is necessarily prepared to deal with all potential probe events, one cannot anticipate all the ways they might be overly simplistic in the nature of the scenario, but because one does not have complete control over the scenario, they have to be able to remain open while also planning for contingencies. Various classes of unanticipated events will happen.
Given CSE’s background knowledge of how conditions tend to unfold in crises, there are multiple cognitive and collaborative vulnerabilities to plan for and learn from. Both planned for and naturally occurring probes must be harnessed, planned for, and exploited.

Next, identifying key pacers and key landmarks serve to set the initial conditions, provide momentum, and shift conditions around events. Pacing, borrowing from the animock usage (which borrows from stage-craft) involves the temporal structure of events ([Roesler et al., 2001]). Pacers and landmarks help orchestrate the flow of activity around the starting and stopping of various events throughout the scenario. As the exercise moves to run-time and analysis, a pacing diagram can be created to represent the relative duration of events and their relationship to actors in the scene. As seen throughout Huachuca, a number of documents were released at fixed intervals that moved the scenario forward independent of current team disposition. Similarly, landmark events such as the commander briefings, set a rhythm for daily operations as well as a natural structure for events. These functions become even more critical given the extended temporal nature of large scale events.

The last two functions are Reflective Forums and Rally Points which respectively serve as ways to externalize and share learning as well as incorporate further contingency planning. Reflective Forums are pauses in the course of events where participants can discuss and exchange thoughts and ideas on the state of events formally or informally. The Reflective Forums serve to both externalize cognitive work and can be coupled with pacers and landmarks to provide a richer common-grounding in the exercise. A final after-action report would be one type of reflective forum. Other prototypical examples would be events such as commander briefings, back-briefs, rehearsals, or press sessions.
These can be integrated relatively passively into the course of events and also provide opportunities for observers to crosscheck participant mutual understanding (Patterson, Woods, Cook, & Render, 2007b). They can also function as Rally Points for the in-situ and meta stakeholder groups to keep the run-time flow on track. Rally Points are likened to instructor interventions which were observed in previous MI research (Trent, Voshell, Grossman, et al., 2007; Grossman et al., 2007) which give planners the ability to make adjustments to scenario events. If in situ or meta observers find that events are disintegrating, the Rally Points serve as reassembly functions so the exercise does not further disintegrate. Given that one does not have complete control, one needs to first say what is built into it. Then there are things you need to be prepared for because various classes of things are supposed to happen. If you have contingencies, you also need rally points. In some sense, because it is large scale, thinking back to Huachuca, if analysts messed up on day one, i.e. if one team is handling an aspect properly but another is missing it, as observers we would want to probe those commander’s briefings further to build a better understanding. Extending that, one could envision the commander as a confederate and now all of a sudden observers can work with commanders to build contingencies into the run. When dealing with more traditional smaller scale events where one only has a half hour for a session, one can not really change, stop, or restart the run. Given these longer multiple day exercises, the organizers and observers have to adapt just like the instructors. There will be no local training value if one simply runs multiple identical sessions as opposed to running sessions contingent on where people are succeeding or failing. This was the power behind the instructor interventions in Huachuca and these functions that show places where instructors intervene is one such way of doing this.
2.8.3 The Run-Time

As the purposes, sequence of events, and learning lab functions become more fleshed out, the run-time starts to solidify. When complete, the script and learning laboratory functions are combined and go into the planning for the actual run-time sequence. This is where plans meet contact with execution. At the run-time we now introduce the different participant groups, observer representatives from the stakeholder groups, observer controllers, as well as participant observers. They all play different roles in the event, and some of those roles will overlap (i.e. the embedded observers, observer controllers).

Figure 2.13: During run-time, participants experience the scenario events along with the learning lab functions as organic elements built into the normal flow of activity.
As the participants and in situ groups encounter the run-time, they experience different elements from the sequence of events. Components from the script are tied to the functions and purposes of why specific groups are there. If a group experiences a probe, observer groups can trace back and explain why the probe is there, what functions it is based on, and more importantly how it is being encountered; this provides explicit traceability. At the end point of the run-time is a debriefing/final reflective forum. Due to the large scale nature of some exercises, there could be multiple reflective forums that define the endpoint. The end of the run-time also constitutes the end of local learning by participants based on their experiences in the exercise. During the run-time, participant groups along with observers and the in situ team take part in the exercise. Participant groups learn locally about the challenges they face as they encounter new injects and components of the staged scenario.

2.8.4 Analysis, Learning, and Understanding

The challenge now becomes how to move from local, to sharable, to general learning. Re-examining the exercise is how we learn. Transitioning above the run-time, we now want to understand what played out relative to the probes, functions, and forums so as to move from local learning to general sharable learning. Across this we are moving from: local learning, where stakeholder and participant adaptation and observation at run-time will often result in usability level findings; toward sharable learning where shared learning between groups and usefulness emerge; and ultimately to generalizable learning in the form of published research, abstracted works, and formal reports sharing generalized understanding. Throughout the learning lab researchers can proceed to utilize process tracing methods (Woods & Hollnagel 2006) in order to begin to
assess these abstract generic patterns of work. Different groups can have different representations that are related to different concepts. The power with the learning lab framework, is now those patterns based on adaptations in the exercise can get linked back to the script. From the script, we can see what probes elicited specific work—what is in the script and about why it was there, what function it has, and what purposes it supports. One probe can link back to multiple purposes and functions. This lets observers and researchers trace back and look for fitment between why they wanted to run the event along with what they had to do to run the event, and finally, related to what actually occurred at the event.

The value of explicit and extensive staging plays out in the analysis and learning part of the framework. As seen in Figure 2.14, the vertical threads mapping test to target extend from principles, to observation, to learning. If one does not know what the probes are then one cannot trace the activity around them. Observers and decision makers need access to all these elements to build the projective map necessary to identify and create multiple typicalities that set up the expectations for what normal practitioners would do. Analysis by distributed observation teams can then note and focus on deviations. In the large scale efforts, this is different because the observation team is distributed, events in one place might trigger looking at distant locations, and the team must be well coordinated across one another to effortlessly shift this sort of mediated observation. The Huachuca case is a perfect lead-up to this. While leaving out some of the scaling complexities, it gives insight to how one could anticipate a large scale observation which depended heavily on this coordinated observational agility and need for cross-cueing. Getting learning out of an exercise in an efficient and timely manner depends on the ability to build in these functions that stage the
conditions of observation coupled with the ability to notice what’s interesting on the fly. This coordinated observational agility, as discussed at the end of the Huachucha analysis, is critical to this. Observers cannot simply pre-code and reduce everything to an observational check list (as with such prescribed exercise methods as those in Chapter 1). What is critical is being able to recognize what’s interesting- and what’s interesting depends on situated context and again being prepared to be surprised. To be able to notice such events, we need to be able to set up events to look for and evaluate how people tend to respond to such situations. We put in specific functions to challenge people, and these functions have a rationale as to why we thought it was challenging. What is interesting is when one finds out that something else entirely is challenging, and this leads to a search to find out what was missing. These are what must get emphasized in the debriefings and reflective forums.

Given this description of a learning laboratory framework from the beginning of planning to the sharing of generalizable learning, elements from the previous figures are combined to create a unified longshot representation of the learning laboratory framework. While this illustrates an overly simplified case, the representation itself provides a starting point from which to guide the next portion of this thesis, a series of critical incident interviews as a means to test and revise the structure and underlying concepts of the framework. This is significantly different than the iterative exercise approach laid out in the first chapter, however it is not inconsistent with what the DHS/FEMA approach to training is, but rather it could help these organizations design these exercises more thoroughly and better. With this initial framework laid out, the learning lab concept is next used to guide exploration with multiple practitioners and researchers to further understand what it really takes to run these sorts of exercises.
Figure 2.14: At the end of an exercise the framework serves as a map of events from which one can process trace and evaluate learning shifts from local, to sharable, to generalizable.
CHAPTER 3

A COLLECTION OF CRITICAL INCIDENTS

To better understand the challenges of large-scale exercise design, a series of interviews using a variation of the critical incident technique were conducted (Flanagan, 1954; Shattuck & Woods, 1994). These interviews provide a snapshot of experiences from reflective researchers and practitioners from communities whose intent is long term sharable learning from such scaled world exercises. Given the scope of the learning lab, there are many potential avenues for discussion, but the goal was to focus on specific studies that had not necessarily been written up in order to explore different ways participants designed studies as well as how they experienced learning break downs and stall points. Semi-structured interviews were conducted with a variety of stakeholders and practitioners from the disaster response community as well as CSE researchers experienced in conducting and observing multi-scale studies and exercises. The participants all represent people who have done, who are doing, and who want to do these sorts of exercises, and the goal for collecting these various critical incident interviews is to identify general patterns across the opportunities and obstacles they have experienced.

The learning lab framework consists of a number of principles that have been extracted through looking at exercises. The experts consulted for the interviews in
this chapter represent years of CSE experience doing this in multiple domains and the question becomes, which of these principles are critical when designing exercises for learning. The details of these critical incident interviews show lots of interesting variability amongst domain experts, but a few key points emerged as a pattern across all the interviews. These interviews serve as data to explore relationships between the many ways CSE investigators have been conducting exercises for years. A number of techniques and methods have been utilized in various forms over this time, but there is still no pre-framework between concepts and exercise design to support building a CTA or CWA on top of. Given that, these interviews also serve as a type of hypothesis test. If the learning lab is a complete theoretically founded approach, then it enables us to compare how different practitioners stage the conditions of observation to watch people at work. Summary tables of learning lab element frequency counts appear at the end of each study to illustrate these trends (see Figure 3.1).
### Figure 3.1: Overview of the frequency count tables and their scatter plot representation for capturing learning lab data.
Across the interviews, the most prominent common themes that emerged in the research stories related to:

- Surprises at run-time
- Adaptations that make you lose track of higher level purposes
- Compromises forced on designers by powerful stakeholders who impede control of the run-time situation
- Planning for observational agility
- Designing decision-difficult challenges
- Maximizing local and sharable learning and imparting general learning

Conducting studies in CSE is grounded in shaping the conditions of observation in natural laboratories through scenario and problem design. Although the field is still relatively young, we as researchers often take advantage of the significant amount of effort and tacit knowledge that goes into creating such events. This chapter serves to collect a variety of perspectives to elucidate some of these techniques used in the field to further refine what it means to use a training or testing exercise as a learning lab. There are a variety of methods we do in general, independent of any formal method. The goal of the interviews was to approach people and elicit reflections separate from the usual methods and descriptions in reported research because we are really trying to extract the elements that experts tend to take for granted as people who do this. We are not only concerned with the literature results of each study, but rather expanding upon what was valuable for the individuals involved in executing the studies, their
struggles and exploits, and ultimately what was necessary to take an exercise and turn it into a learning study.

The learning lab concept began as an attempt to generate reflection on the processes behind getting valuable observations and lessons from field exercises. A significant amount of tacit knowledge lies behind good scenario design and the ability to simulate events that move beyond textbook cases (Woods & Patterson, 2001). Similarly, organizers must extend a significant amount of effort to balance practical relevance with scientific credibility while monitoring and coordinating the observations themselves. Exercises cannot be simply designed to succeed skewing the vital learning and feedback designers and engineers may need. As mentioned earlier, in other simulations there is often a lack of failure conditions so as not to erode practitioner confidence in real world transfer. Purely technological approaches to design tend to specify engineering needs based on use cases with specific goals for given exercises while the critical cognitive requirements of the joint cognitive systems are often underestimated. The learning lab framework emphasizes the interplay of such challenges as they are built into envisioned worlds and takes a problem based approach to provide greater understanding of the systems at work. Given that technology fundamentally changes the nature of work, this also creates new demands in order to be able to test and capture the dynamics and adaptations of domain practitioners in realistic simulations. These are the skills and practices these researchers and practitioners have had to cope with and develop over the years.

A number of candidate interview participants were chosen and these individuals were then emailed a brief problem statement along with a short attached document outlining general scalability challenges in exercise design. To avoid confounding
interview sessions, no formal explanation of the learning lab concept appeared in
the read-ahead information. For the interviewer, questions were anchored around
the learning lab framework to guide the lines of questioning and identify common
trends and challenges across various exercise experiences. In each interview session,
the framework was used by the interviewer to guide knowledge elicitation questions in
order to evaluate various aspects of exercises and then reflect on the learning point
of view to get back and identify the local, sharable, and generalizable learning. The
initial framework served as a roadmap, which the interviewer could mark while taking
notes in each interview. Analysis from the interviews shed further light on the sorts
of dynamics between components within the framework as well as general revisions
across the multiple stages.

3.1 Critical Incident Interviews

The interviews followed the general technique laid out by Flanagan (1954) and
this chapter uses that same organization to relate and interpret the findings.

<table>
<thead>
<tr>
<th>Flanagan’s Critical Incident Technique</th>
</tr>
</thead>
<tbody>
<tr>
<td>1. Determine the general aim.</td>
</tr>
<tr>
<td>2. Develop plans and specification for</td>
</tr>
<tr>
<td>collecting data.</td>
</tr>
<tr>
<td>3. Collect the data.</td>
</tr>
<tr>
<td>4. Analyze the data.</td>
</tr>
<tr>
<td>5. Interpret and report the findings.</td>
</tr>
</tbody>
</table>

Figure 3.2: Flanagan’s original critical incident method.
3.1.1 Participants

Twelve participants expressed interest and volunteered for the critical incident interviews. Ten of these interviews served as the cases outlined in this chapter. The two other interviews, conducted with military training and Homeland Security practitioners, encompassed more general lines of questioning and while useful interviews, were not examined using the framework. The ten participants consisted of stakeholders, organizers, researchers, and practitioners with a broad history in exercise design, observation, and participation. Interview participants represented a corpus of fieldwork in a variety of mission critical domains including nuclear power, emergency response, and defense training. Basic background information was collected for each participant prior to each interview.

3.1.2 Materials

Each interview participant received an email with a two page outline attachment describing the purpose of the study and areas of interest. In the email, a short resilience motivation was tailored for each domain and sent to representative stakeholders and practitioners. The two page outline (see Appendix ??) presented a brief overview of new challenges to exercise design and learning due to scaling. These topic areas framed the initial questioning in the knowledge elicitation sessions.

3.1.3 Procedure– Collecting the Data

All invited participants received an email containing an outline of the research goals. Interviews were arranged and conducted in person or by phone and lasted between 30 minutes and one hour. First, the research goals outlined in the read-ahead were briefly reviewed with participants. Participants were then asked to recall personal exercise
and research experiences that had components of scaling outlined in the read-ahead material.

The semi-structured interviews were conducted in two stages, similar to typical CDM style [Klein, Calderwood, & MacGregor 1989]. These began with personal introductions and a repetition of the research goals. In the first stage of the session, interviewees were asked to share a recent personal experience and then in general regard challenges faced based on their own exercise and study experiences. This first stage served to simply get the story-lines out. The interviewer took detailed notes using the learning lab framework to annotate participant experience as learning labs. The second stage of questioning then revolved around using the annotated framework to go back and revisit specific components of the story that either fit in or did not fit in with the framework, or that were genuinely interesting and warranted further investigation. Across these stages, common questions that emerged tended to revolve around asking:

- What knocked them off plan and how were missed opportunities mitigated?
- What opportunities merged to take advantage of?
- What obstacles or new challenges emerged?
- How did you anticipate and adapt to surprise events?
- How did hypothetical demands get inserted to probe cognitive work?
- How were ceiling and floor effects anticipated?
- How do you share what you think you’re learning
Interviews ranged in duration between 35 minutes and 1 hour and 30 minutes. All notes were taken by hand and audio was recorded when possible under participant permission. Two of the interviews utilized a third stage which incorporated a tangible prototype of the learning lab framework for the participants to “do, say, and make” (Sanders, 2002) filling in and naturally linking learning lab stages and components into a navigable structure.

3.2 Results

The resulting transcripts and artifacts were assembled into a series of cases organized by common domains. When possible, multiple perspectives from the same case were captured to provide further context and insight. The findings were analyzed, reported, and interpreted based on their domains and how they represented learning lab approaches. Summary tables of learning lab elements are provided at the end of each case. Using the defined components from the learning lab framework, these tables represent frequency counts of observed categories from each description.

3.2.1 Case 1: The New Designer

D.D.W.- Research lead, Research & Development Center, Westinghouse

“The first time I did this I got thrown into what was supposedly an evaluation of new technology. In 1980, EPRI was engaged upon nuclear reactor control room redesigns with Westinghouse. This was one of the first studies of operator performance in emergencies. We had a variety of purposes, and a number of stakeholders- vendors, regulators, manufacturers, operators. Everyone was a stakeholder because there was not a lot of data about what operators do in emergencies. I was appointed the research lead to evaluate and design this safety parameter display system for future nuclear control rooms. I came in after Westinghouse got the project, and from one point of view- this was going to be a way to evaluate future designs. It was for future use and the formal title was it was an evaluation.
The plan was to do an evaluation of new display designs—safety parameter display system for nuclear control rooms—to help people focus on safety. In the end, however most of the learning wasn’t on that, but in general on what’s difficult in handling emergency situations in control room activity. They had a high-level plan, but that high-level plan that was inadequately specified to actually run the exercises let alone how to analyze them. It was too high level. It was specific enough that someone put out an Request for Proposals, we won, but it was under-specified. This gave me the degrees of freedom to design the study, to design some hedges into the study, and to design some focusing mechanisms, to build stuff into the exercise to make it more likely to find observable and interesting results.”

This case comes from the dawn of CSE and illustrates elemental challenges to the conduct of future system human evaluation studies as well as a glimpse at the beginnings of driving CSE concepts behind scaled world observation techniques. What was initially designed as an evaluation study of new display concepts turned into seminal research on the cognitive difficulties inherent in anomaly response. This was a future system, and envisioning was built in. It was observed that just by watching practitioners interact with their new designs a significant amount of local learning, especially in the form of usability was immediately relevant leading to rapid re-tuning of designs. However, what proved quite challenging in this case was sharing that information in a more general sense. Before the first study was even run, the research lead knew immediately there was inadequate planning for contingencies. The exercise was originally planned as a canonical path exercise. The significant initial challenge to getting effective learning out of this, it turned out, depended heavily on how much the exercise set up ahead of time. This theme is evident across almost all of the cases discussed in this thesis. It was obvious to the researcher that their team could not simply perform a two factor design comparison because this was an entirely new kind of system. Introducing new technology fundamentally changes the nature of the work,
therefore focus must be placed on what about that work is difficult. In this case, the PI had the deferred authority to redirect and redesign the problems built into the case. While he did not have control to redefine the study as a whole, he did have the ability to make the scenario difficult (in stark contrast with the Strong Angel where no one on the research side, observer or participant, had deferred authority).

As we shift from planning to analysis, the value (and advent of) process tracing (Woods, 1993) became a critical component to be able to map out the exercise and cope with a massive data overload from the raw simulator data when it came time for evaluation. Given the nature of the exercise, the researcher did not have strong themes and general patterns of cognitive work to relate to coming in, the themes had to be found in the data. Lots of events were identifiable but there was no coherent integration. In order to get around being caught in this inductive trap, a trap as Dekker and Hollnagel (2006) describe in accident investigation as the “What You Look For Is What You Find” principle, the researcher had to develop a way to try to discover these larger patterns on the fly. This entailed moving away from context specificity to concept specific understanding.

Tying back into the learning lab framework, there was co-location between both observers and practitioner teams in this case. The simulator rooms were designed with observation functions built in and this facilitated non-obtrusive viewing. For the simulation environment itself, the run-time for the scenario was a formal infrastructure, a simulator based on mathematical models of a powerplant. Observers did not have to worry about the run-time breaking down aside from the simulator itself breaking-the computer micro-world infrastructure was generally robust to breakdown. As we start to regard more open loop scenarios where the run-time is not as fixed, such as
with the three field observations, recuperating from a scenario or run-time break down creates a break in reality that can be tough to recover from. The researchers were then able to build probe events were built in to the scenarios and given the strong infrastructure, researchers did not have to worry about doing things that don’t turn out the same way. Again, as exercises scale and this same level of control dissipates, this emphasizes the functional need for ‘rally points’. It is impossible to be able to anticipate all the post conditions or reverberations around events. This will increase the difficulties around logistic factors as other surprises accumulate. Due to this, there need to be ways to re-assemble on the fly to make the event work. The consensus from the interviews was that by seeding ‘enough’ complexity into the system, the opportunity for variations will emerge.

Figure 3.3: Categories observed in the New Designer case.
3.2.2 Case 2: Designing the Right Sized Hole

W.C.E.- Nuclear Technology Systems Division, Westinghouse

Coupled with the previous case, ‘the right sized hole’ case provides additional insight into the issues faced when designing scenario challenges, especially regarding how to anticipate floor and ceiling effects. Planning to design the right cognitive problems to serve as probes in first of a kind system is not easy, and as the previous case refers to, when left to engineers intimately associated with the system, they often miss the point entirely leading to operators learning the wrong lessons. The basic approach related with this case can be seen as what we would now view as an edge based or decision centered testing approach (Rousseau et al., 2005). From a fundamental CSE perspective, researchers need to know where to pressure and know where to push in a scenario and in both the planning and the analysis of a learning lab, the investigator needs to be able to insert and uncover those cognitive pressure points (Woods & Dekker, 2000).

“For our simulated worlds- we ended up picking out the right sized hole which in a sense looked like a trivial problem- but it turned out to be decision difficult. It was decision difficult as soon as you look at the way the span of the various systems laid out, and the range of automated control- it was a silent automated system with a capacity limit- the edge to stress is right at the edge where auto and Manuel have to deal with each other. The decision at each end is trivial- I’ve got nothing to do, theres a leak and some radiation, so pump like crazy. The hard part was right in the center where the operator and auto had to deal with this edge effect. The experts were less than enthused with our right sized hole because it was a small hole. They were being nice and giving us simulator time. We were on the back shift when the simulator was empty, they hung around for a few extra hours to help us out. The licensed operators were the guys who we were reporting the results for. We were really trying to validate whether these functional anomaly messages gave you better insight and better performance than the standard annunciator based limit crossing stuff. We didn’t know what event it was ahead of time- we made up the
event on the fly. We were then able to do a simple comparison between
the current systems with the same physical event happening each time,
because it was all canned. All the set-up parameters became saved and
it became a defined event on the training simulator, so they record it as a
“Dave and Bill special No. 1”. You recall that, they reset the whole
plant so you’re running at 100 percent power flat and level, you tell the
guy in the booth to throw a hole in there, and its exactly the same hole,
in exactly the same spot, with exactly the same thermodynamic impact.
That event, that hole size never occurred to anybody. They had been
licensing operators on holes and on pipe breaks, for decades. The expert
assessment was if you got the extremes covered, then by definition you’ve
covered the maneuver space in between. That’s a flawed assumption! A
big one. It seems to make sense, if you can do tiny holes you can do big
holes. But it’s not like that at all if you look at the cognitive work and
the decisions you have to make.”

The original scenarios designed by the instructors and engineers for the nuclear
control room simulation had some inherent flaws. They decided that the way to test
the system using a Loss of Coolant Accident (LOCA) was to present the extremes of a
failure case. They assumed that by testing the system with a massive pipe leak and then
testing it with a minor coolant leak, they would have a robust system test that spans
all the conditions between the two ends. What the researchers discovered, however,
was that the real decision difficult cognitive work was not when failure conditions
were at the extremes, but rather when leaks accrued slowly in between those extremes.

How could the instructors have discovered this was the case? One approach would
be to simulate or game the conditions and stochastically and empirically throw game
states at operators. This is very much the experimental psychologist model where
the researcher would randomly pick system states and if one picks enough of them,
they might arrive at some interesting cases. But one has to consider the complexity of
real-world operations. In a nuclear plant (or even simulator), one could never be able
to pick enough independent variable system states in this manner. As the number of
variables in play go up, the degrees of freedom go off the chart and as touched upon in
the first case, interpreting the empirical results becomes even more difficult. Even if
an exponential number of cases were run, sometimes hard sometimes easy, one is still
left with needing to explain it because it is all post hoc. The fundamental question
that has to be answered is what about the case was it that made it hard? This puts the
researcher on the spot, they need to be able to reverse engineer and have the deductive
skill to take the end result and deduce the cause. This is a fundamental difference
between feeling around in the dark scientific exploration versus a “I’m an engineer
that has to test a system, I can only do it once, so I have to do it right” mindset.
There should be hypotheses going in based on designed problems that are supposed
to focus on validating or disproving your hypotheses. The key, however, is that these
must remain flexible and open to interpretation. If one only goes in and expects there
is data in there somehow and somewhere, and you find lots of little things this means
there was no hypothesis going in.

As Roth et al. (1993) observed, there are many classes of cognitive challenges that
could be incorporated into a scenario, and there is no prescribed systematic way to
employ their use. If a key probe/challenge/independent variable in the scenario proves
too easy (ceiling) or too hard (floor) when employed in a pilot study, a number of
alternative strategies can be utilized. To reiterate the sentiment that opened this
case, cognitively complex scenarios must be designed that effectively pressure the
joint cognitive system such that coordination complexity emerges. When the goal
is to probe suspected boundary conditions and cognitive pressure points (Woods
and Dekker, 1999), CSE researchers need to be able to elicit and gauge practitioner
performance that is complex, but neither impossible or too simple. In Staged World
studies, it is critical for researchers to be able to design problems and instantiate them in face valid and meaningful scenarios for practitioners (Woods and Hollnagel, 2006). As researchers design the situations, probes, and context for these problems, they can assess how JCS performance responds. If performance (the dependent variable) is found to be completely floored or ceilinged, this is indicative of bad probe design (time to manipulate the independent variable differently). If these effects are encountered, researchers must re-evaluate what interesting situation points they are looking at, and must also explicitly work through the target test mapping of the probes used in detail (Woods, 2003). For specific examples based on the nature of the problem and probe design, researchers can take a number of steps to remedy floor or ceiling effects. First, consulting with experts and reflective practitioners to make sure the conditions you are simulating are in fact challenging is a good strategy. Morel, Amalberti, and Cahuvin (2008) found this out recently when studying sea-fishing skippers. Their simulation required a gradual worsening in weather conditions leading to safety limits. To mitigate floor effects they consulted with expert skippers to determine these levels rather than do it themselves. Another strategy is seen in Case 5 outlined below, where the authors presented a variety of scenarios to experts and the case they chose and the nature of intelligence analysis allowed them to raise and lower the level of difficulty of questions analysts faced in case the scenario proved too easy or too hard. In Patterson, Rogers, and Render (2004), floor effects were observed when very challenging probes were employed which profoundly highlighted how poor the system operated under stressful conditions. In that case, this serves as a performance baseline that redesigns can be measured against. Another option, as suggested by Miller et al. (2006) is to turn to the practitioners themselves and utilize the results from the pilot study
as a source for critiquing so as to generate group discussion and debate with the practitioners to give further insight in how to re-calibrate the study.

Like any of these design problems, honing in on sufficiently complex scenarios must be part of a bootstrapping process (akin to Potter et al. 2005 for CTAs). Each step in the scenario design builds upon the previous one. As one gains a better understanding of the pressures in a domain, they will further appreciate what characteristics to incorporate. If a scenario is just too easy after that first pilot study, this could be a good sign that the scenario captures normal routine activity well. Complementing the strategies listed before, CTA strategies can also be employed to help refine the complexity of scenarios to move beyond routine and better reflect complicating factors in this envisioned world phase. Potter et al (2005) suggest the use of storyboard walkthroughs, participatory design, WOZ, and rapid prototype techniques to refine the initial model of work. Researchers then may discover that event tempo or uncertainty for example are the critical issues to be pressed harder in the scenario. Finding ‘the right sized hole’ was not difficult for these researchers to pick up on because they were able to seamlessly coordinate perspectives between the equipment, the instructors, and the operators to understand why how cognitive challenges develop. These scenarios and the probes in them become prototypes for discovery. Pilot studies represent initial hypotheses, and these prototypes with scenarios can employ a number of strategies to refine and better capture the demands of cognitive work.
3.2.3 Case 3: My Three Types of Exercises

E.M.R. Research Lead, Science & Technology Center, Westinghouse

The following case illustrates the general range of staged worlds CSE researchers operate in and highlights a number of specific difficulties faced from planning to analysis across a number of nuclear control room simulations.

“I’ve primarily done large simulations, mostly been nuclear power. I’ve done other work, there’s also the issue of how to create scenarios to support more evaluation of other kinds of systems- which are at least as challenging but in other ways. One obviously stands out... where the issue was to try to craft scenarios that were cognitive challenging to put people in these difficult situations. It came out of the general 3-Mile work where the notion was, the event that actually happens often differs dramatically from the simple canonical ways people think about. So there was real interest in understanding how operators are likely to respond to complex situations that deviate from procedures. A second one also interesting to the NRC- we were trying to understand how new technologies change the
ways crews perform—also done for the NRC... we didn’t get to manipulate the scenarios we got to observe the scenarios and interview the participants. That was interesting in itself, because that’s more opportunistic. More recently, it was two nuclear studies but the real issue was new systems that had been put in place. So you need to define a range of scenarios that will allow you to assure yourself operators will be able to handle a range of situations in real life. Those are therefore broader and more programatic in the range of situations you’re trying to model—frequent events, high risk events, cognitively challenging events—they’re all challenging in different ways. One is you’re trying to broadly sample the space of situations. Two, you’re trying to opportunistically almost use critical incident techniques to derive interesting things from the situation, and the third one, you’re trying to craft situations that will reveal particular behaviors.”

The predominant themes that emerged in this session revolved around techniques to build expertise, the importance of face validity, as well as the logistics challenges of evaluating new systems. In the first two types of exercises, where researchers are broadly sampling a space of situations and where they are opportunistically deriving situations of interest, the biggest constraint to cope with was time. Lack of time affects the researchers’ ability to build up familiarity, a similar initial challenge in Huachuca, and represents a basic challenge to generalist human factors approaches. It is critical to be able to build up domain knowledge to be able to identify problems and disruption when it comes time to observation. Fortunately by the time of these studies, observation and research resources in nuclear simulators had become more mature, instrumented, and largely designed for compared to the first two cases.

“What we did is we observed a subset of the scenarios. We waited until later in in the 2 week period until they were well-along in their training, and we observed a couple of the scenarios. I think that we alternated doing observations. We observed them, and we then had a chance to interview the operators both after every scenario if we had some questions and at the very end asking them general issues of concern. In terms of observation, this is where I could tell you the meta level of what we do. In terms of observation, the issue is to look for cases that are unexpected. Cases
that deviate from the expectations of the developers of the system. So in particular, I looked for cases where procedures broke down. In nuclear power plant situations, observation is really straightforward. You have instructor control rooms, two-way mirrors—observing is not hard. What is hard, and varies from situation to situation, is the domain expertise that is required to understand and evaluate what you are observing. But for these studies, because we designed them ahead of time, we really understood the events, and we had the procedures in front of us. We knew the expected behavior, we knew when actions were taken and not taken. The other thing that makes nuclear observation easy, is operators articulate when they’re going to take an action.”

This reconfirms that as with the first two cases identifying meaningful disruptions is still a difficult task. Those disruptions serve as cues to ask questions during operator interviews. Temporal constraints affected this and had to be worked around. The researchers went into each session with detailed checklists to guide interviews, but often the interview times were significantly limited (i.e., planned for two hours, afforded 30 minutes). The researchers needed to quickly be able to integrate disruptions from observation into the targeted interviews. For the seasoned investigators, this was effortless but they recognized this fluent opportunistic integration is not easy. Methods to support this type of observational agility are critical.

In these first two types of exercises, research was piggybacked with training and interestingly enough, research results were also incorporated as part of the training. This is a good example of built-in reflective forum. At the end of the exercise, the researchers wanted to provide a general reflection for all involved instructors, practitioners, and designers. This was built in so that the researchers could explain their observation to the operators as a way to reflect on the cognitive topics while specifically reviewing the alarm system operators were training on. At times the stakeholders (utilities) would ask for direct implications on training, but given their
fundamental adherence to procedural training, the researchers could not suggest outright that findings might suggest people be allowed to override the procedures. This highlights the need to cope with and maintain stakeholder goal alignment throughout to avoid such conflicts.

In the third type of exercise where researchers craft situations to reveal particular behaviors to evaluate new systems, some very difficult scaling challenges arose. Two recent exercises of this type were designed but they were not piggy-backed on training. Initial planning proved to be decision heavy when designing evaluation studies for these new systems. Researchers need to decide upon staging a wide range of situations and resource constraint judgments will start to accrue. There are limitations to how many crews can be run, how many scenarios can be run, and how much training one can provide on the new system itself, all of which make it difficult to run the cognitively complex situations a researcher might ideally want.

“When you’re trying to evaluate a system, you’re really trying to define a space of situations to turn through. You’re not necessarily trying to choose a hard one. You’re trying to sample the broad space. You need to consider things that are cognitive hard- as well as things that are physically hard, and things temporally. You have to think about all those dimensions. As well as boring task analysis and the range of situations people are likely to be involved in- and sample those. Have to think about what are likely to be the most critical in safety space- in power plants its the risk critical.”

Defining this multidimensional space is not straightforward (the learning laboratory framework being one such attempt at representation), and the added scale of the system, without being tied to training incurs escalating financial costs. Very little support exists to aid planning and help in the ability to choose just what types of scenarios or scenario challenges to be utilized in such cases. Being able to cater
scenarios given constraints on training time, how much time is spent on training versus how much time is spent on design testing are both difficult tradeoffs to weigh. However, this must be balanced before a researcher can begin to evaluate and design the best scenarios that satisfy these multiple dimensions.

Figure 3.5: Categories observed in the ‘My Three Types of Exercises’ case.
3.2.4 Case 4: The Dayton EOC Tornado Exercise

L.G.M. Research Lead, University of Dayton Research Institute
E.S.P. Expert Observer, C/S/E/L, The Ohio State University

In the following disaster response case, scenario design planning was complicated at the last minute by face validity issues when reviewed with practitioners. Researchers were able to adapt for the scenario planning shortcomings and quickly re-plan and develop a mature and successful study due in strong part to the quality of the observation team. Interviews were conducted in this case with both the research lead and one of the expert observers from the study. The study was conducted with the Dayton EOC with research and stakeholder goals interested in studying how people from different organizations not used to working together prepare for infrequent events. Specific interest was focused on looking at logistics issues and how computer supported collaborative work (CSCW) could support similar operations.

“We just usually tag onto training exercises. Unusual to create your own because its labor intensive and don’t usually have enough domain knowledge to do that well. A representative came down to meet with us and talked to us about different types of exercises. They have a distinction between tabletop exercises in which they don’t actually play- key players come in and they talk through what would happen- vs. an exercise they actually play. So we definitely wanted one where they wanted to play. So we found a bunch of stuff online, and that identified a company that sold us an exercise. It was pretty cool. It was a really fancy excel spreadsheet and it had exercise injects. You could enter who the players in your exercise were, and it would take those and meld that into the exercise so it could be tailored to specific players. This was crazy, spent a ton of times working that- going through the spreadsheet- then the day before the exercise- the guy who runs the county EOC finally reviewed it, and said this wouldn’t work. Meanwhile we had invited all the representatives from the county, 23 people, so we were down there until the early hours of the morning redoing it by hand. His biggest complaint, completely legitimate, wish knew about it earlier, had really strong information push. The white cell
would just call in with information periodically throughout. This does not happen, you’re there in the EOC and have to track stuff down. He wanted to change it so we’d have knowledgeable people in the white cell- so last minute we recruited people from Dayton Police Department, UD Campus Police, really knowledgeable folks and some regular grad students- that could be coached. And it was a great innovation, made the exercise much more realistic.”

At time of planning, the researchers did not know what cognitive challenges to expect going into the study. A popular software package being marketed to EOCs and emergency response groups across the country was chosen for the general scenario creation. In attempt to make the scenario realistic and challenging, the PI worked with a highly experienced subject matter expert with a background in tornado disaster response to create a series of injects for the exercise.

“These were all really tough decisions along the way but the planning was really hard. We were kinda in over our heads, didn’t know what we were doing- Tracy [the SME] knew a lot so we kind relied on her so we were just kinda relying on her.”

The software then incorporated their inputs and automatically generated a linear course of events for the exercise run-time essentially creating a pre-time stamped MSEL. The goal of keeping a linear canonical chain of events was to attempt to avoid scenario stalls in case people became stuck given this was the first time the research team had run such an event. With the run-time planned after a month of development, the PI sat down with the EOC director to discuss the exercise the day prior to execution.

“I met with the EOC Director and his first comment was, this wouldn’t work.”

Sitting down with the EOC director, significant shortcomings in the scenario were identified and these were largely due to the very simplistic software engine used for
scenario design. As the director quickly determined, the biggest problem with the software model behind the scenario creation was that it was only based on information push to the participants. As seen first hand in Huachuca as well as observations at FDNY’s Fire Department Operation Center (FDOC) and New Jersey’s Regional Operations and Intelligence Center (ROIC), and in just about any real response, these situations are never push only and depend largely on the actions of the individuals involved and their responses to get new information- it is largely information pull. There is a significant context-bound relationship between push and pull which has to be planned for, that was completely absent in the software model. The software also had predetermined settings for disaster scales: low, medium, and high. The only difference between the three settings was the number of push events put into the scenario. There was no insight or observability into the actual nature of the complexity of those challenges in the system nor did they explicitly map to a defined scaling dimension. These challenges resulted in the PI spending 8 hours redesigning the scenario with the Director by hand and rebuilding the exercise for run-time. They augmented the event shortcomings not through more software, but by introducing a highly adaptive white cell to handle the key probes and pacers the software created, coupled with a robust observation team of HF and CSE professionals. A group of graduate students, originally planned to be observers, were instead incorporated into the white cell. This group did not have a lot of response domain experience so the PI complemented them with local practitioners from emergency services that could help the team think on their feet as run-time challenges occurred. By working together they filled in serious gaps from the simplistic software and could readily handle a number of other software limitations, as well as support a much more realistic flow of
events at run-time. The other significant contributor to the exercise’s success was its observational support design, as described by E.S.P., one of the expert observers in the observation:

“The PI was not observing, she was the coordinator of the observers. That role was really important as she was able to real time re-script it because phone calls didn’t happen when they were supposed to. The software based the calls in and phone calls out as the events- decided to have the same people do phone calls in as phone calls out. Interesting because a lot of phone calls in didn’t happen at times they weren’t supposed to. Every time they hung up, it would ring again. They couldn’t dial. They had 3 big tables, about 30 people in one room. And however many of them calling out, they were simulating it all and nobody could call in.”

The entire exercise was labor intensive and highlights the key support roles needed. The PI’s goal was to put her most experienced people as data collectors and observers and then support them and the exercise run-time with the adaptive white cell. Due to the scenario’s brittleness, the software designed script of events was thrown out early in the exercise because real-time events did not map well to its assumed time intervals. This resulted in the PI planning the prescribed injects on the fly based on observed activity and cueing the observers what to expect when. Much like in Huachuca, the observation team was coordinated with the PI, who also interacted with the white cell in tandem. A significant amount of activity was occurring in a number of places, and this coordination proved crucial to its operation. The exercise concluded with an After Action Review that encouraged a significant amount of positive participation from operational participants leading to multiple examples of local learning (i.e. usability- phones could not be used because they were too far from desks) and sharable learning (i.e. new agencies were identified for coordination to cope with problems never encountered before). The observation was then analyzed
from a cognitive perspective and synthesized into an exploration of information flow and coordination which was published (Militello, Patterson, Bowman, & Wears, 2007) and later briefed to EOC stakeholders.

3.2.5 Case 5: Ariane 501 Intelligence Analysis CTA

E.S.P. Research Lead, C/S/E/L, The Ohio State University
D.D.W. Advisor, C/S/E/L, The Ohio State University

The following case looks at two perspectives from a successful CTA highlighting a scenario re-design early in the planning stage of the study. The researchers in this study were funded by stakeholders who were interested in learning about data overload in intelligence analysis. The researchers incorporated stakeholder groups from AFRL Human Effectiveness along with NASIC to design this CTA, and upon reviewing a
moderately developed candidate scenario with stakeholders, the researchers met some challenges early in the project.

“We had funding to do a study that illustrates from beginning to end what a Cognitive Task Analysis is. We started with Zaire, wanted to run a scenario- and we knew we wanted to use the NASIC analysts. The problem with Zaire, there’s no air force there. We talked to NASIC, via a guy who retired that served as an SME. We have this Zaire thing worked out, we had played with it for 3-6 months, put together all these news reports to create a scenario with the documents and all that. We took NYT articles and all that. 80 percent of those were just copy and pastes out of wires. One comment was, the stuff in the media is always wrong, and there’s no air force component. We met the analysts, and out of nowhere Dave said, Ariane 501. I asked him later how he came up with that, because we had never talked about it.” - E.S.P.

This represents an interesting twist on the typical Law of Fluency (Woods & Hollnagel 2006) and highlights one of the consistent themes across all these cases: run-time scenarios are a balance of control with real world complexity. The immediate question the PI (and most likely the stakeholders too) had at the time was why this particular accident case would be any more relevant, than the scenario first proposed. This suggestion was based on expectations about the class of event the rocket accident entailed. What became evident as the scenario, materials, and apparatus for the study came together was that there was so much variety in the real world regarding how accidents are reported on, assessed, and analyzed, that only capturing a fraction of this real world complexity would still lead to a significant amount of insightful findings and challenges for exercise participants to work with. Unlike in the nuclear cases or Kings Mall, there was no training system to piggyback onto in this case. Instead, familiarity with the domain was supported by working with trusted insiders with the stakeholder groups, actual analysts, and communicating with system designers.
“So we were focused on, what was the theme, how was it face valid to the scenario. We first tried to do more of an intel scenario. Rwanda, crisis, humanitarian, different groups, how do you build up a database. What was different in this from the aviation ones, we weren’t walking into a well honed training system—there was no training system. A lot of the issue was how do you build up the simulation, rather than adjust cases from the instructor set, we had to create one. We had to create the infrastructure to present it to people and build documents into a database. We took a shortcut, I said don’t worry, we understand the nature of the problem because we know something about the event. It was an accident, and we knew the diagnosis was different than normal—it was a automation software accident. Now what’s interesting about the database relative to understanding the event, will come out as we analyze what they did… to understand the event even though we know its really about understanding the event. We know all this background and we understand accidents. In that sense, it was a safe shortcut.” - D.D.W.

With the scenario designed, the researchers were able to leverage the design into an existing intelligence analysis software system under development. Like with the nuclear simulations, this created a relatively stable run-time environment for each participant. A follow-up study conducted by Miller (Miller, Patterson, & Woods 2006) used a similar set-up, but was able to piggyback on training for a different perspective of method and data collection. In the first study, there was a high degree of self-selection of analysts because people were interested in learning the new system as well as socio-cultural circumstances (again reinforcing the Francophone ‘practitioners are stakeholders’ tradition and that we are not simply detached observers). The analysts who volunteered understood that a lot of the technology changes developers were throwing at them were perceived by the developers as helping the analysts have a more user centered perspective. Analysts, however, wanted to have a voice in this. In the latter study, people would participate as part of training. There was a workload gap and a large part of the training was built around mentoring. This created an ideal
situations to further study how to bring novices analysts up to speed in professional settings and parallels the value of mentorship as a learning strategy as seen earlier in the military and fire approaches. The scenario and problem set was essentially the same between the two studies, but now the data was how the senior person evaluated and gained an understanding of what the novice was doing. Both examples illustrate the significant amount of practical, local, and generalizable learning generated from the two studies. These were both published from a cognitive systems perspective in the literature and continue to serve as landmark CTA cases for the intelligence domain (Patterson, Roth, & Woods, 2001; Miller et al., 2006). Beyond this and anecdotally, this study served as a powerful illustration of what it means to share learning beyond just a training context which came in the final briefing after the first study (a reflective forum):

“The main challenge was then getting it into their site. That was Janet Miller’s idea—there’s an unclassified training facility that has pathfinder installed! She was a grad student taking classes at the same time. She was going to do a dissertation in the same area. The SME was a pathfinder champion, he wanted it to be used, he had pathfinder brown-bags once a month. He sat there while every participant ran with me. A few chairs down, he just wanted to see what the software could be used for, what was good for. We could take snapshots, we could stop everything, we had a complete amount of control. Sharing the results also went really well. Had software developers and analysts that participated in the study. We presented the stuff.... software developers said that’s not what analysts do. An analyst stood up and said that’s exactly what we do, you have to listen to them!”

The true power of good CSE, and ideally what learning labs should be able to deliver, is the last line in that quote. The results from the exercise went beyond just the local design and training context the NASIC stakeholders were interested in and the researchers were able to capture and relate the abstract research conclusions of
the investigators and be able to tie them back to workload and software development challenges. In this reflective forum, the investigators were able to use the exercise study as a communication path between the actual professionals and the system developers to highlight the profound cognitive challenges in their work. This complementarity, being able to use well planned, well executed, and well analyzed research to communicate and meet multiple purposes from training, to practitioner reflection, to system design perfectly illustrates the successful transitions from research to design insight to generalizable learning and work domain understanding.

<table>
<thead>
<tr>
<th>Staging</th>
<th>envisioning</th>
<th>observation</th>
<th>goal alignment</th>
</tr>
</thead>
<tbody>
<tr>
<td>APRIL Researchers</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>Practitioners</td>
<td></td>
<td></td>
<td></td>
</tr>
<tr>
<td>NASIC Developers</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Run-Time</th>
<th>infrastructure formal</th>
<th>informal</th>
<th>Probes</th>
<th>Pacers</th>
<th>Rally Points</th>
<th>Reflective Forums</th>
</tr>
</thead>
<tbody>
<tr>
<td>computer sim face valid scenario</td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

<table>
<thead>
<tr>
<th>Learning</th>
<th>Local Usability</th>
<th>Shareable Usefulness</th>
<th>Generalizable Understanding</th>
</tr>
</thead>
<tbody>
<tr>
<td>raw data design data Findings Briefing Full CTA CSE Papers</td>
<td></td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Figure 3.7: Categories observed in the ‘Ariane 5 Intelligence Analysis CTA’ case.
3.2.6 Case 6: The Failed NASA Aviation Study

D.D.W. Research Stakeholder, C/S/E/L, The Ohio State University

The following case is interesting because it represents a boundary case. It details a study on the cognitive effects of long-haul fatigue on aircraft pilots that never made it out of the planning stage. The failed NASA aviation case highlights just how difficult it is to orchestrate stakeholders and focus enough goal alignment momentum to transition closer to a runnable exercise. As this work also shows, however, just because a failed study never made it to staging, that does not mean that there are no avenues for learning. Experiences and themes that arose from pursuing this work went on to become generalizable as seminal CSE research on design process and evaluation study approaches (Woods & Sarter [1993]).

“We were trying to do a fatigue study with Curt Graber. There’s a fatigue side, there’s a cognitive side, perfect piggyback. To assess fatigue, you should look at these cognitive effects. On the other hand we were trying to understand airplanes, then fatigue becomes a way for us to reveal problems coordinating with automation. It seemed natural. We worked on a scenario, we worked with instructors- we did a much more top down plan in terms of learning themes and elements- it was explicit. Information integration, you need something that people have to integrate. You can forget to bring forward, you notice something early in flight, its now relevant later in the flight- it was related to how auto affects cognition. Tunnel vision- what kind of events and probes in sequence would challenge and tell you if whether or not you are tunneling on key indicators. We knew we needed probes and all that before, but that’s the first time it became explicit.”

Undertaken roughly a decade after the nuclear studies described in Case 1, this represented a more mature scaled world design approach. The CSE researchers had a number of high level cognitive themes they wanted to investigate, they saw this as an opportunity to explore new coordination challenges with increasingly complex
automated aviation systems, and they knew how to build in complexities to the script and scenario to elicit these challenges. They had a face valid scenario grounded in the complexities of the domain, and there was top down structure to undertake the investigation with complementing training. Unfortunately, the CSE was not the only stakeholder at the onset of planning, and the multi-stakeholder balance became even more challenging as the true scale of the study began to take shape.

“But what happened, the problem with the fatigue study, you had too many groups to coordinate. Too many purposes to coordinate. Too many organizations to provide a valuable purpose that could contribute resources. You have the commissioning people. The difference was with long-haul fatigue, crews had to fly with them to collect fatigue data. The logistic cost of observation, do you have observation teams that can spend the same amount of team as the performance/operator teams? Operator teams are flying all around the Pacific, and some research needs to fly around with them. What’s your N? 20 teams, 20 flight crews, 10 on front end and 10 on back? [You had]... to look at the difference between the same simulated event people at the beginning of the tour and the end. Notice the cost issues- airlines won’t come back and say we’ll take 2 hours of their flight duty time by giving you an hour and you can do it within- or one scenario with practice. We were designing the scenario. We did two iterations. We had a good scenario, every bought in. We had worked with NASA social psych team. The issue was, could NASA management pull off with the airline industry, and specific airlines and pilot unions to cooperate. What’s interesting there, in many ways its a large scale exercise. The aborted fatigue study illustrates some aspects of what it means to be large scale. We had to coordinate multiple parties at run time. But it just died under its own weight. There may have been something specific, but underneath it was the realizing it was just too hard. You go to all this effort then you lose cases when there were delays off duty. It was just too brittle. If it broke down, there were just too many consequences for the participating groups. The airline would lose, the pilots who volunteered would lose, and the researchers would lose. It might have worked if someone had a lots of money to invest, if the surprise was bigger, if the urgency was bigger, if there had been an accident attributed to fatigue...”
As related to the interviewer, the study shared many of the dimensions of large-scale exercises. The temporal and spatial scales alone, mixing real-world flight observation (both passive and invasive) with simulator studies presented significant logistical challenges in just planning how to get participants involved. These would incur huge financial burdens and given the legalities of pilot flight hours, it was never guaranteed participants would be available after multiple legs of flight. The researchers would have to coordinate multiple research groups and multiple phases of each trial in this distributed manner and as related by the interviewee, this ultimately crushed it under its own weight. Money was an object, and unlike the aftermath of Three-Mile Island where high level stakeholder goals were formulated around redesign at all costs, aviation did not have a similar catastrophe or bellwether events that warranted the same funding or tenacity- this wasn’t driven by failure. One could cynically see this missed opportunity as another result of typical bad design/bad outcome cycles which only look to CSE and HFE after a significant problem leaving us relegated to sweeping up at the rear of the parade (Woods & Tinapple, 1993). However, one can also see this as a missed opportunity due to the scaling challenges that exacerbated research planning coordination. In looking at the 25 years of large-scale ‘collaboratory’ failures, Bos (2007) imparts that “creating affiliation between scientists is easy, but creating larger organizational structures is much more difficult.” Continually grounding diverse groups of specialized expertise, especially expertise that is often tacit (as indicated earlier), is difficult to share over distance- in such knowledge management situations, distance still matters (Olson & Olson, 2000). Similarly, when collaborating with multiple diverse groups, cross-institutional boundaries often percolate, as seen here, where goal conflicts and legal conflicts become significant hurdles as logistical challenges
are met. There is no magical support technology or easy solution to such situations, and it goes to further emphasize the need for continuously grounded goal alignment within scopes of controllable scales.

Figure 3.8: Categories observed in the ‘Failed NASA Aviation Study’ case.
3.2.7 Case 7: US Army Approach to Division Training

W.C.E., Lt. Colonel US Army, (Ret)

The following case illustrates the modern US Army approach to Division level intelligence and combat training. These large scale division level exercises utilize distributed teams of participants and observers, take place over multiple days and across large areas of land, and are designed to both evaluate and cultivate learning. In Chapter 2, the limits of spatial scaling were mentioned regarding Kings Mall referring to a physical upper-bound on what could be designed for staging. Interesting parallels can be made with how the US Army has come to cope with these issues when they hit a similar physical wall of complexity (and scale) over 50 years ago. As the United States prepared to enter World War 2, the Army faced significant scaling and complexity challenges where new equipment, doctrine, and tactics had never been battle tested. The response to this was to stage series of large scale exercises before going into battle that came to be known as the Louisiana Maneuvers (see Figure 3.9). Using approximately 400,000 troops, 3400 square miles, and costing $20.6 million (in 1941 dollars), one of the largest scale war games ever was held across Louisiana and North Carolina to test maneuvers and gain insight to thoroughly test doctrine, command, and logistics before going into battle (Jarynowycz 2001; Barbier 2003).

Historical analysis and research on the Louisiana Maneuvers is rife with illustrations of the challenges faced by analysts, participants, and stakeholders from the event but looking forward at the bigger picture, Army training realized that it could not hold such events if it were to train and develop effectively. Gathering all the equipment, resources, and personnel together to test high level command strategy and tactics was not a sustainable option. With advances in computer simulation, the Louisiana
Maneuver moniker has been revived to apply to modern Battle Lab initiatives which use mixed simulations to train battle staffs, new technology, and new doctrine without the significant resource burdens of physical deployment (Wilson, 1996). Approaches to military training have evolved rapidly in the past 15 years (Salas et al., 1995; Shattuck & Woods, 2000) as well, and to a degree have been incorporating more cognitive type challenges.

“The Army has set up Leavenworth as the battle command training program. Its goal was, give division and corp commanders, 2 and 3 star guys, guys that command 10,000-100,000 people- give them first battle experience before the first bullet is volleyed. To where these 2 star division commanders feel they have engaged the enemy. That experience should feel so real they couldn’t tell it apart. Basically, your division field headquarters is rebuilt. Drive it in, trucks and ramps go here and there, whatever your policies are. You build it, they wrap the war around you. Hotline corp commander, a white cell answers the phone. If the intel side needs a guy
to talk to its, there. They surround you with simulation and white cells so you can’t see light.”

Until relatively recently, Army training programs have largely been simply scaled up versions of existing approaches to individual training focusing on skill acquisition and as such, have generally been behaviorally focused with a significant lack in the training research regarding distributed teams (Salas et al., 1995). As with aviation and the evolution of CRM described in Chapter 1, aviation realized that only focusing on individual skills did not automatically lead to higher level coordination and operations in the cockpit. In response to the accidental Vincennes shootdown in 1988, significant military research efforts began to focus on distributed team and decision making approaches to translate into training systems which has resulted in a number of cross-training and mental model training approaches today.

“It starts back to the LOCA- you can’t walk in and cut the biggest pipe in the plant for real, you just can’t. Your ability to create these conditions in a physical world is limited. You’re never gonna have 2000 walking wounded in city. You’re ability to do things in a simulated world is way higher. That is what the Army has learned. That layer of decision making needs to sit above a layer of simulation on the bottom. At the bottom, machines that run and do the math- and then they find there are things that machines can’t simulate. So you have the bolt-on white cells. They are the broker between the unexpected phone calls, somebody with judgement makes stuff up to fill in the gaps. The flexibility is in the white cell staying in the spirit of things. The trick is making the exercise credible, interesting, and a learning exercise.”

Typically, planning in these sorts of exercises starts with looking at training objectives. From the overall objectives, individual tasks and goals are identified. These can be further broken down into individual level and collective level tasks to determine at what unit level and then what individual tasks should be stressed. As repeated with each participant and trainer interviewed, if one is just evaluating
command and control, then one should not go down to the individual level. At this point training teams start to build the scenario employing story-boarding techniques while drafting a ‘road to war’ to develop the narrative and story-line. This is used to establish the broad goals and then subordinate orders will come from that. While research and specific training initiatives have emphasized the value of more decision centered grounding, it is often not specifically planned around or developed at this critical phase of exercise planning. These essential task lists rarely have a cognitive work context, and when they do, it is often implicit as a result of the work-task design practice. These work-task designs are often based around producing artifacts and targeting this area is a prime CSE/learning lab leverage point to provide decision making contexts around those artifacts. Shifting back to current planning methods—after a MSEL has been constructed, individual events are designed in and placed in the exercise to make it successful. At the same time a transition from operational to exercise planning occurs as a parallel development process. On the simulation side, trainers then select the appropriate town, cities, locations, terrain databases, and decide what type of enemy to use.

“Brigade and division- 250 people being trained/evaluated. White cells maybe half that. Observer controllers, you have lots of them. [They] have to be in places 24x7. Three shifts of them running around, 1 per organization: one guy grading the aviation, one for a division staff team of 15 people, 25 division guys, 10 OCs... If you’re fighting a good fight, you’re in it until done. We had one mission which was a defend. If you fought it well, we were in a deliberate high intensity fight for 3 days. OCs go through and rewind and replay what happened. They put marks in the event. They talk with each other, go through things, they’re always coordinating with each other. Ear buds, all wired in. They have their whole hierarchy too. They have the big dog observer and specialists all running down. They have little team meetings, what is our focus today, what do we think will happen. Because they are neutrals, they come in and say what’s your plan of attack, how are you going to be attacking- they
get a sense, then they talk to the red guy- how are you going to attack them? Its like watching the world series of poker- you see the hands unfold across. They have that omniscient view, so they need to know where to be and what to watch for. That’s always critical on the evaluation side.

Then there’s also the red team. They play by their own rules, and they live in a very similar world as the blue team, the guy fighting. They are fighting in a doctrinally realistic setting, surrounded by the same battle simulators. If this guy moves his tank regiment forward, other guy fights it. Stratego goes on. Rules of the red guy, he has to be a fair and honest scout team quarterback. Supposed to play as the adversary. Allowed to be creative, but only in the sense as long as you’re replicating ‘typical’ creativity of what the guy could do. You can’t do French Foreign Legion tactics if you’re a Soviet Regimental Commander.”

We can trace many parallels to the learning lab approach and the military planning approach, but there are multiple unique challenges and lessons learned that can influence better learning lab design. The most significant lesson learned is similar to that from the EOC exercise, and that is the profound value of having a flexible and well developed white cell (as a meta-group) to create that ‘master roadmap’. Especially in the military context, while local learning in the form of skill training accompanies experience, the predominant effort is put into learning from the AAR. As discussed in Chapter 1, to have a successful AAR as a type of reflective forum, the scale and effort put into the functional role of the white cell is a critical component. For sharable learning, the AAR is critical for continuous improvement and continued learning. Despite the maturity of modern simulation systems, significant observation and white cell support is required to both keep the exercise on track while essentially process tracing the event from as many perspectives as possible. Exercisers constantly operate under the goal of reducing the time from observation, to AAR, to exercise implementation. It is the researcher’s responsibility then to be able to turn such findings into generalizable sharable research.
Figure 3.10: Categories observed in the ‘US Army Approach to Division Training’ case.
3.2.8 Case 8: Observation Support in the OR

S.G., Research Lead, University of Virginia

The following case is noteworthy because it captures not just a CSE observation of real-world multi-agent laparoscopic cholecystectomy procedures, but it also relates the story of a customized software development effort to build an automated data collection and editing tool for observation analysis. The researchers in this study had to not only learn-up the details and become familiar with the nature of the medical procedures, but they also had to develop advanced observation software so they could monitor the multi-agent interdependent medical staff richly and from multiple perspectives in situ.

"Everything we’ve done with scenarios has been really hard... Clearly when you’re observing a team, that’s really hard. And we were trying to measure team performance in the operating room. There are 8 people in the room, and its very difficult to figure out- and we didn’t really know exactly ahead of time what to look for. There were some certain things, we wanted to check for, i.e. did the second surgeon brief the team prior to the incision. But beyond that, we wanted to measure team communication and coordination. This was a gall bladder surgery setting. We picked that one because it’s easily at most an hour long procedure, it’s fairly routine enough to get good data, UVA performs an average 1 a day or 4 a week. It’s not perceived as a very high risk surgery. The procedure itself can take up to an hour."

The high-level goal of this body of research was to look at team communication in healthcare. While the gall-bladder procedure observed here is not typically risky, these were real-world surgeries and to that degree encompass a number of potential uncertainties and time pressures. While temporal and physical scales are compressed in this case, respectively roughly 1 hour operations in a single OR, a limited number of researchers had to richly observe up to 8 members of a highly fluent interdependent team. In each surgery, one or two observers had to track multiple finite events by each
individual surgery team member. At times this could range from transcribing direct verbal transcriptions of utterances and communications, to viewing and annotating detailed surgery procedures from video monitors. The observation teams needed a way to meaningfully integrate these diverse information feeds and assemble them in a coherent organized manner to analyze post hoc. Further complicating this fact was the notion that these were real-world operations. There was no planned sequence of events, so observers had to have the observational agility to adapt and re-orient to events as they unfolded in order to try to get insight into and capture this distributed tightly coupled technical work (Cook, Woods, & Miller 1998).

“At the minimum, there’s an attending surgeon, resident surgeon, drug tech, anesthesiologist, circulating nurse. At a minimum there’s 5 people, but there could be more... so we’re trying to figure out- the idea was if we could give an intervention like a check list- i.e. make sure you checked all the instruments before start, make sure gave antibiotics. As opposed to if you trained them in team communication and coordination skills, call outs, briefing the team, are you going to improve performances. At first, we really weren’t sure how to measure performance- what are you measuring? One thing people do, is count utterances- who’s talking when. We started with that, and just write it on a piece of paper, and that’s hard. What ended up happening, all this different kinds of data we wanted to collect- and we ended up with 40 different pieces of paper, and you can’t coordinate all that. Plus you have the video tape. So just putting the data together in a common data set is really difficult. And that’s how it all evolved... we developed this RATE tool.”

Existing observation tools such as MacShapa could not interface with the video hardware streamed from the OR equipment and commercial software such as Observer could not handle the scale of multiple mixed video and audio inputs needed. Similar to Huachuca, the researchers had to improvise. Different from Huachuca, these researchers developed their own team-behavior tracking tool, RATE (Remote Analysis of Team Environments) from the ground up. They were able to overcome the technical
hurdles of essentially instrumenting the ORs and medical staffs relatively passively and
combined this with rather clever software to synchronize incoming feeds with observer
note-taking. Observers agreed ahead of time in each operation what events to code,
and these were selectable options in the software to fulfill. What became interesting,
however, was the ability to go beyond just the pre-coded events and presenting the
ability to flag and annotate periods of activity ‘on the fly’ as processes unfolded that
were interesting to the observer. As seen in Huachuca, such flexibility is imperative to
build in at observation time to then synchronize and revisit in analysis.

“So we created this software package that lets you combine 4 videos,
and this is really hard because you have to time-stamp and synchronize
them all. It took us all a really long time to technically figure out how to
do it. They’re all on different computers with different start times. So we
built a cart with 4 computers, and video capture cards. Each cart had 1
monitor, one mouse, and a switch for all the computers. And we made
it portable. It all ran off an Access database and that’s the beauty, it’s
all time stamped. When we loaded the video later, start encoding it, pull
time again. It worked great. You could then jump anywhere [based off
your notes]...”

This last element captures the true power of such a high level human-indexed
observation artifact that this software approach enables. The raw data feeds that make
up the process trace are no longer just a collection of time-stamped and synchronized
playback files, but rather they become a higher level navigation mechanism for the
playback of events themselves. Analytic context is provided around certain events,
but these can be revisited and re-analyzed given new insights or different analytic
approaches. While this could potentially shorten the time between analysis and
sharing of learning, such playback methods could be highly valuable when used in
conjunction with AARs or reflective forums with the practitioners themselves. This
is a classic learning technique well known to sports fans, but also widely used in
usability and interaction studies (Jordan & Henderson, 1995). Usability practitioners often assemble video highlight reels of user frustrations to relate the real challenges to developers. Similarly, in urban fire response, Omodei (2005) assembled wireless video cameras inside firefighter helmets and then used a human factors interview protocol as a cued-recall technique to develop rich traces around events. The seemingly simple technique of being able to repeatedly replay a sequence of events and interactions from multiple perspectives is profoundly powerful. One could envision such techniques being combined with similar software as developed for the OR to support organizing multiple agent traces. The resultant rich process-oriented performance data can then be collected to explore new ways of supporting learning by coupling mentorship through playback elicitation, critiquing, and feedback.

Figure 3.11: Categories observed in the ‘Observation Support in the OR’ case.
3.3 A Need for Planning Support

Looking at the common principles that come up in more stories and more interviews, the summary tables indicate that exercise staging plays a critical deeply connected role that guides what researchers planned to capture and how they could look into multiple ways of analyzing events. Across these stories we see this in the form of:

- Surprises at run-time
- Adaptations that make you lose track of higher level purposes
- Compromises forced on design by powerful stakeholders who impede control of the run-time situation
- Planning for observational agility
- Designing decision-difficult challenges
- Maximizing local and sharable learning and imparting general learning

Significant challenges come out of these trends which create a difficult tradeoff space to navigate when planning. As exercises scale up, they become more expensive and more time consuming. When the researcher has no control then they need a big picture to work with to explore potentially valuable areas. However, when they do have some control or if they are designing the whole event from scratch, they have to make decisions based on what to include and how to maximize learning in the exercise. Generally what CSE research tends to focus on is telling people about the analysis side and results from studies. Support tools such as MacShapa, RATE, and Observer have been developed to make this analysis process more efficient and amplify the
ability of researchers to record, capture, and do basic analysis on what was captured. The synthesis and higher level findings however are still left to the insight of the investigator to pursue. Few of these tools support this much more difficult staging side of exercise design. We did see one state-of-the-art scenario design tool used in the Dayton EOC case which revealed to be quite brittle when it came to run-time. Similarly, we can look at the Army example and observe the significant scale of human resources dedicated to simply the planning and observation necessary for a large scale AAR.

There is a profound need for planning support, and to respond to that need begins by developing a roadmap of the problem space- this begins with a longshot. Designers need a way to capture and store all the elements and work-arounds observed in this Chapter while trying to work through a planning study as these worlds scale up. Because the scale of control and the scale of the exercise in the staging and running has become so large, it is no longer feasible to rely on just one person to integrate these things. Each of the cases in this Chapter, barring the Army exercise, was effectively tied back to one or two key individuals responsible for the exercise design and execution and keeping the plan under control. In all of these, one or two people tended to keep the momentum going, mitigated surprises, and when issues came out usually only one person acted as the center of gravity. In all of these, as well as the field observations from Chapter 2, there was usually only one or two people that served this role. However, given the challenges of scaling up, we cannot get away with this anymore given the increased scale of space, time, and multiple teams. We need to be able to represent and support this staging phase just as much as the run-time and observation capture phases. The learning lab framework enables researchers to focus
on supporting how we are doing the staging and how the staging links into the higher
level phases of the learning lab.

As designers and staged world observers, we want to be in the position so we can
control the run-time situation. The only parts we can watch and capture data to
analyze is what we control. We need to be able to explore the potential range of
control at run-time and capture so we can have data to look at. Facets of what we
need to control in order to abstract the cognitive work that goes into staged/scaled
world study design and evaluation into a representation interface design. By making
these elements explicit, as was seen in the “do say make” exercises for two of the
interviews, by giving people a navigable representation to organize the material around,
we present an open-ended design tool that makes things that are often inconsistent
and conflicting, and overall genuinely difficult, an external artifact. The overall goal
then is to implement these functions into a sharable design seed in order to elicit
discussion and propose new decision support tools to help effectively control and create
a run-time where behavior and context is capturable so that researchers can perform
and share effective analysis that contributes to multiple levels of learning.
CHAPTER 4

A LEARNING LABORATORY PLANNING SUPPORT TOOL DESIGN SEED

This chapter provides specification for a prototype design seed of the learning laboratory planning tool as a Decision Support System (DSS)- the “Learning Lab Designer”. This chapter contains design concepts and descriptions to outline the software development of the application. This chapter represents a transition from analysis to concrete design following a similar applied cognitive systems engineering practice for decision support system design as described by Elm et al. (2003). The design seed is based on the notion that there are fundamental patterns of cognitive work to support in the domain of staged/scaled world planning, conduct, and analysis. The demands imposed on researchers and planners, especially as these exercises scale in complexity, can compromise successful learning yield. This is a challenging problem compounded by some of the existing brittle support tools discussed in Chapter 3. These tools tend to either focus on just the scenario design aspect or just on the data collection and organization phases of a study. As discussed previously these software tools often attempt to simplify and trivialize these functions for the sake of convenience, and in doing so often further constrain design and evaluation. The driving motivation behind the learning laboratory as a software planning tool is to
enable researchers, stakeholders, and planners to be able to go deeper and richer into the complexities that arise from planning to learning - while everyone developing such tools tend to think in the opposite development direction. The goal here is to link the broader set of CSE derived functions with support for planning.

Based on the conceptual grounding in CSE principles, the initial framework established in Chapter 2 was reviewed and re-evaluated based on the critical support requirements for staging which emerged from the interviews with practitioners. The results from the meta-analysis combined with data from the critical incident interviews were translated into specific functional requirements to guide the design and conduct of a learning laboratory support tool that:

- Focuses on staging
- Represents planning to make it explicit
- Uses a longshot to represent the entire exercise

The analysis and interpretation of the results from the previous chapters serve as the basis for decision support concepts framing the design of a Learning Lab design planning tool. The fundamental representation for the planning tool is built on a longshot (Woods & Watts 1997). Longshots for navigation have proven immensely rich in representation design which adhere to the basic CSE representation design principle the navigation mechanism should be a model of the topic to being navigated. The framework serves as a model of the topic for support. The rest of this Chapter outlines the requirements for the software that frame the Learning Lab Designer prototype. The walkthrough will be more informal and is meant to read similarly to...
a typical user guide. First, an overview of the system is provided and then example wireframes are used to illustrate:

- Interactive playback of previous learning laboratories
- Interactive visual planning and modification for exercise design

4.0.1 Scope

The learning laboratory support tool is designed to aid in both the playback and planning of exercises. The design prototype consists of three main work environments to be described in detail below:

- Learning Lab Project Chooser
- Playback Viewer
  - Annotated Project Outline
  - Learning Lab Overview Map
- Learning Lab Planner
  - Project Overview Map
  - Inspector Pane
  - Component/Function Browser

4.0.2 Analysis and Design

Before describing the interface, we must first describe the revised learning laboratory framework as a functional system which specifies both the key elements, goals, and their relationships. From this model of the system, the functions, goals, and decisions
can begin to be developed for the interface design. Results from the critical incident interviews are used at this phase to guide and provide context around the design components. To convey this analysis, the revised learning laboratory framework is broken down hierarchically into three stages—staging, run-time, and learning—and functionally assessed within each stage. From this functional breakdown, design is then focused on supporting these relationships, on multiple levels of abstraction, resulting in the software windows and panels that constitute the user-interface of the planning tool application.

Figure 4.1: The learning laboratory progression from outline to output.

As first illustrated in the framework in Chapter 2, as stakeholders move along the diagram over time vertically, we can see the different stakeholder group perspectives. Planning is shown in relation to a series of events being run in time. The timeline of planning and learning is positioned relative to the run-time axis. Moving horizontally along the run-time, the relationship to the perspectives of different stakeholder groups is shown throughout planning and exercise envisioning. The multiple stakeholder groups have a variety of purposes—some might be technology oriented, some might want to
demonstrate a system (again, plurality), some are research oriented. Because there can be multiple learning objectives from these groups, it increases the need for all stakeholders to be involved in initial preparation and all stages of envisioning.

This learning lab framework represents a functional support model for exercise planning. This resulting design artifact can be used to either trace back previous exercises or stage new ones. The previous portion of the chapter specified the work that must be supported by the decision support tool and now serves as a model of what needs to be displayed in a support interface. The rest of this chapter describes the design seed based on these work and presentation concepts to describe how the learning lab will be displayed along with descriptions of system interactions within it.

4.1 Envisioned System Description

The goal of the envisioned Learning Lab Designer prototype is to create an integrated application using multiple interacting visualizations to support researchers and exercise planners both review exercises, plan new exercises, and explore the analysis of exercises. Before the user can start planning an exercise, they need a longshot of the end to end process. Between the exercise observations and the SME interviews, it becomes clear that there is a large amount of variation between how exercises are conducted and utilized. However, from planning to analysis there is a common workflow behind learning laboratory designs. The three basic phases to design a learning laboratory are described as:

- Step 1- Planning an Exercise. This is where multiple stakeholder goals, purposes, and interests are conceived, scripted into a scenario, and supported with
multiple learning laboratory functions. This entails resource costs, budgeting, and logistical planning.

- Step 2- Executing the Run-Time. This is where participant groups engage in the scenario and are observed by researchers using a variety of observation techniques and methods.

- Step 3- Analyzing the Exercise. Analysis is the first step toward abstracting learning from local, to more sharable and generalizable forms. This is where the true stories of the exercise play out and new understanding of cognitive work can be developed.
4.1.1 Opening the Application

The Learning Lab Designer, in its application prototype form, is first opened by clicking on its icon in the Dock or by double-clicking its icon in the Finder. The web-based version requires a single click on the application’s launch button.

Figure 4.2: The Learning Lab Designer being launched in an envisioned workspace.
The **Design Launcher Window** then displays the three main tasks the user can choose to perform: creating a new exercise, viewing an exercise, or viewing a video tutorial on Staged World Exercise design narrated by Prof. David D. Woods. Prototype and wireframe screens will be reviewed for each functional option.

Figure 4.3: Design Launcher Window: from here the user can choose whether to plan a new exercise, view an existing exercise, or explore the tutorial assistance.
4.1.2 Tutorial Library

The Learning Lab Designer contains an extensive video tutorial library to guide the user throughout the conceptual CSE grounding of the learning lab, as well as provide examples and ‘screen-casts’ of working with the interface itself. Topic areas are organized based on the learning lab stages and are provided in a scrolled list pane on the left hand side of the window.

Figure 4.4: The Learning Laboratory Video Tutorial Library allows the user to explore the CSE concepts and application interactions.
By either selecting a topic from this list by single-clicking on it, or by typing into the search box in the window title, the user will see corresponding help topics appear in the right pane of the window as they navigate the help database. Selecting a topic will immediately expand its corresponding content to present a video and short text description of the topic. When the user is done exploring the Tutorial Library, closing the window will immediately re-open the Design Launcher Window.

Figure 4.5: The Learning Laboratory Tutorial playing a video clip of Dr. Woods explaining key envisioning concepts for the planning of a learning lab.
4.1.3 Viewing an Exercise

Project Chooser Window

Every exercise created in the Learning Lab Designer exists as its own XML document. As a user edits or builds a new learning lab and new elements are added to an exercise, they can be saved and then opened and viewed from different machines. Each learning lab available on the local machine is displayed as an icon within a scrollable window that the user can select.

Figure 4.6: All available exercises are represented by a high resolution icon and brief title.
A scrolled list of domains is provided to narrow in on specific areas of interest. From the drop down menu, exercises can be further sorted by domain or by CSE patterns. Because learning lab filetypes are indexable, the search box at the bottom of the pane can be used to target even deeper content such as researchers involved, keywords, or stakeholder identification. After identifying which learning lab the user wishes to explore, select it from the chooser pane on the right. To open the learning lab, either double click on its icon or click once to highlight it, and then click the ‘Choose Exercise’ button. This will then replace this window with the Playback View Window.

Figure 4.7: The user has chosen to view Emergency Response learning labs and has selected the Kings Mall observation.
Playback Viewer Window

After selecting an exercise, the document will open in the Playback Viewer Window. The Learning Lab can be ‘played back’ using the video transport controller in the top left of the window, or the user can select individual phases of the learning lab to play, pause, or explore. Clicking and dragging within the main window repositions the contents in the window. The user can zoom in and zoom out from the learning lab to explore different sections in detail.

Figure 4.8: The Playback Viewer enables the user to explore the learning laboratory from a variety of perspectives to examine different stages of the exercise, how relationships played out, and how run-time challenges mapped to initial goals and learning.

180
Clicking on elements within the learning lab provides additional context in the two support panes. To the right, a floating inspector pane contains basic exercise information, as well as basic data for the stakeholder groups, participants, and researchers. As the user clicks on different elements in the learning lab, the corresponding data will populate this pane. On the bottom of the screen is the annotation data for the learning lab. Each available perspective is provided and time synced to the playhead. In the case illustrated here, the raw observation data, the exercise designer, and one of the Observer Controllers has annotated the current phase of playback to provide their perspectives about 3/4 of the way through the run-time.
Key artifacts from the learning lab are hyperlinked from within the document. For example, citations made in the annotations or papers published as generalizable research can be linked into the learning lab’s database structure and opened in a separate window. When the user is done viewing the exercise, they can either quit the application or close the main window to return to the **Design Launcher Window**.

Figure 4.9: Artifacts such as this Human Factors paper can be linked to the learning lab document and then be opened on the fly.
4.1.4 Planning an Exercise
Learning Lab Designer Window

This last section serves as a walkthrough of the exercise planning designer itself. As mentioned in Chapter 1, to ground this part of the tutorial it is best to instantiate with a simple real world case. For this walkthrough, we are going to look at the basic planning functions for conducting a hypothetical canine Search and Rescue (SAR) training event. SAR-Dog and working dog training is unique because it shares an interesting theoretical place between HFE human performance approaches and more behaviorist animal training traditions. Recent research has looked at using human performance methodologies to study how human:dog teams are trained with regard to expertise and task vigilance (Helton, Begoske, Pastel, & Tan, 2007). From a CSE perspective, some researchers have looked at human:canine coordination as a competency model for future technological system design (Murphy et al., 2007).

In the current threat environment canine teams have seen increased use (in often cases as a response to brittle technologies) in hostile check-zone operations in Iraq (Mott, 2003), search and rescue in remote areas (FEMA, 2003), and de-mining and explosives detection operations (Helton et al., 2007). In spite of their increased usage and long training history, there are no national training or competency guidelines for human:dog performance evaluations in these fields in this country. To this degree, volunteer and professional groups alike are constantly looking toward new training methods so they can better coordinate in larger scale mixed operations. Therefore we can identify three groups of stakeholders with different learning goals and a shared interest in the future of human:canine search operations: CSE researchers, practitioner handlers and search team managers, and response organization representatives. Also,
these operations reflect scaling challenges. Conducting area searches for lost hikers or disoriented children, for example, is a typical response operation that highlights a number of large scale challenges on top of these general trends. These types of operations take place over large physical scales (i.e. multiple miles), can last for multiple days, and involve the coordination of distributed teams. These provide the initial purposes for conducting a new learning lab.

We will start with a blank exercise that we can start populating with these stakeholder groups. The learning lab designer contains all the functions necessary to map out initial planning conditions which can then be adjusted across subsequent stages of planning and analysis. When the user first opens a new learning lab planner, they are presented with the basic three stage structure of the framework.

![Learning Lab Designer](image)

Figure 4.10: Starting a new plan in the main Learning Lab Designer window
The user begins the planning by selecting the ‘Roles’ tab in the Learning Lab function browser at the bottom of the screen. Icons representing stakeholder groups and participants can then be dragged into the main window, the learning lab stage. As each icon is added onto the stage, fields will be automatically created in the the Inspector pane on the right. After all stakeholders have been added and all fields have been filled out with initial stakeholder goals and group information, the stakeholders can coordinate to align their goals and identify complementary opportunities before they can go on to develop the scenario.

Figure 4.11: The user places the various stakeholder groups onto the learning lab ‘stage’ and fills out their initial goals prior to scenario generation.
As the three stakeholders convene over the course of a couple weeks, individuals bring new goals and purposes to discuss. These are added across the horizontal section of the ‘Purposes and Functions’ portion of the stage. In our hypothesized case, the State Police representative would like to use a training opportunity to evaluate the readiness of a local volunteer team. Citing concerns that Summer was approaching and historically her own search team had been overwhelmed by children and nursing home patients getting lost in an expansive nearby wetland preserve, the State Police representative wanted to evaluate a local volunteer team’s ability to support State operations.

The volunteer rescue group saw this as an opportunity to test out access to new mapping and communications technologies and decided to explore these new functionalities for this event. The CSE representative also saw this as an opportunity to look at distributed planning between the two groups, as well as a way to evaluate the potential of using HRI metrics to assess handler:canine team performance in the field. As this initial planning coordination continues, these additional goals and purposes are added to the stage in a similar interaction as adding the initial Roles icons. After milestones have been marked and all initial goals added, stakeholders can begin to define the participants in the ‘in-situ’ group as well as participants for the event. In a similar fashion as before, icons are dragged from the function browser and placed in their respective locations. With the actors and observers selected, and goal alignment converged, planning transitions into developing the details to the approach for including learning lab functions and scenario development.

With participant groups and stakeholders set, much like in the building block approach it is now time to develop a narrative, scenario script, and identify areas to
Figure 4.12: As stakeholders work with one another to incorporate multiple goals and purposes, the actors for the event can be identified and included in the staging.

insert learning lab functions to support exercise learning opportunities. As stated, the State Police representative expressed concern over conducting area searches in large wooded areas. Stakeholders decide to simulate two young children getting lost on a Summer evening in a two square mile section of wetlands. Helicopter and infrared searches are usually unable to operate well in such areas given the vegetation and abundance of animal life, often necessitating canine responses. This cluttered and large spatial area also poses ample challenges for communications equipment. The stakeholders use this basic story to serve as the narrative for the scenario. The narrative sets the stage for where events take place, the nature of the emergency, the
time of day, a general sequence of events, and other potential emergency factors to consider. The user clicks on the 'Script Development' button at the top of the main stage to isolate the next section of planning to receive focus. First, the user drags the ‘narrative’ icon from the function browser and can either double-click or use the inspective to fill out the basic narrative information.

![Image of the software interface](image)

**Figure 4.13:** Stakeholders start the staging of the scenario by creating the exercise narrative. After the narrative is established, the sequence of events and learning lab functions are developed in parallel.

From here, stakeholders create an event list (MSEL) and add learning laboratory functions in parallel. The event descriptions are derived from the narrative and are
place in a general temporal order. So as not to be too scenario driven, planners can immediately go back and view exercises purposes and explicitly mark-up what purposes map to what learning laboratory functions and what functions are instantiated in the events. For example, if the CSE researcher wished to create a probe to gauge how well a handler:canine team could re-synchronize their search efforts, a communications breakdown or GPS failure could be inserted. Observation could then be planned to focus on how the agents adapted their own efforts, and what work-arounds were pursued to regain grounding with the larger search effort. To do this, the user simply drags a line between the three events, and they can then tag the test:target mapping path with this hypothesized demand. This same interaction can be used to fully label the sequence of events.

Figure 4.14: As the sequence of events are laid out, they can be visually linked to the learning lab functions as well as initial goals and purposes to provide a hypothesized roadmap of the scenario.
From here, the stakeholders can then begin to execute the exercises and embark upon resource acquisition and logistics prior to the event itself. The planning support tool provides this initial roadmap for constructing the event, planning for observation, and creating contingencies, to prepare for the run-time. At this point, the support tool functions similarly to existing observational analysis tools to facilitate conducting detailed process traces from the exercise observation. By capturing the evolution of planning as well as explicitly preserving test:target mappings, the planning tool represents a unique support artifact from which to explore further and uncover complementarities, mismatches, and surprises in the actual field observations.

4.2 Using Design Seed Prototypes for Collaborative Envisioning

This walk-through is not exhaustive and a number of interface elements and interactions are not fully developed in the prototype. The tool serves as a design seed and as a means to collaborate with practitioners and researchers to determine what would be useful toward developing innovative new approaches to staged and scaled world study design (Woods & Hollnagel 2006).

The Learning Lab Designer was created to integrate collaborative envisioning into exercise planning across all stages of design. It is a tool for discovery and can support collaborative envisioning in for R&D, exercise planning, and scenario design. The planning and scenario design stages are treated as a collaborative envisioning process up front (Schoenwald et al. 2005). These scenarios help designers and practitioners share and examine how to seize upon aspects and organize the exercise events around the actual flow of domain knowledge and around challenges and disruptions. The mappings between these goals and purposes that translate into the events are explicitly
maintained. The scenario and exercise planning design process itself serves as an envisioning tool that both describes and can be used to share the complex interactions (and potential unanticipated side effects), stories, and adaptations in work from which to initiate feedback from practitioners, designers, and technologists. From a CSE perspective, this type of collaborative envisioning helps us navigate the entire learning lab so as not to get lost in the details and quickly transition between abstract and concrete elements in its navigation.
As intoned by Woods (1993), research and practice are mutually reinforcing and to be successful, CSE researchers must look to field settings as natural laboratories for long term learning. Developing new scenarios, new simulator technologies, and new artifacts as tools for discovery must be complemented by resilient staged world design to bridge between application and research. Accelerating the growth of expertise touches upon the three dimensions of learning, scaling, and resilience. Observers must try to practice synchronization while practitioners in situ are also often practicing synchronization. Going into these observations, multiple stakeholders from researchers to instructors to technologists are concerned with learning about what contributes to efficient coordination and synchronization as well as how to reflect, share, and facilitate that for all groups involved. Instructors and planning committees run these events locally and pragmatically, but these local and pragmatic efforts can be shaped and expanded to gain a greater understanding about the cognitive work processes and therefore better facilitate the local pragmatic growth of expertise, improvements in coordination, and enhanced resilience.

Many of these exercises are conducted in the context where the long term learning goal is anticipating and accommodating new technology and new skills and how to
take advantage of them. This requires an understanding of the work and frames our responsibility for staging exercises as learning laboratories. The goal of staging exercises as learning laboratories is to steer learning opportunities in such a way as to get useful learning for all involved parties and ultimately facilitate insight into new understanding about the nature of cognitive work in the domain.

5.1 Summary of Contributions

The meta-analysis and interviews in this thesis provide a framework and design seed of a planning system for running exercises as learning laboratories. The learning laboratory approach, at its core, is designed such that exercise designers are prepared to be surprised. The learning laboratory approach is based around planning to be resilient so that designers, observers, and problem holders are prepared and ready to take advantage of unplanned and surprise events. On top of this conceptual grounding, this work as a prototype planning tool is complementary to efforts such as MacShapa, Remote, and Observer in order to build specific tools that support the collection and integration of data records in a manner that supports efficient and focused analysis so as to abstract sharable and generalizable learning.

- Contingency planning is needed because disruptions will occur, preparing for the realities and complexities of run-time when dealing with hundreds of people for multiple days means you need to do contingency planning.

Across the meta-analysis and interviews, it is evident that pre-planning and contingency planning are critical to envisioning how disruptions may occur and anticipating successful ways to recover. As was seen in Strong Angel 3, not being able to recover from run-time breakdowns can be disastrous to a large-scale exercise and
nullify even the potential for learning on any level. The fire department instructors at Kings Mall were able to use observer controllers to coordinate and “brute force” exercises back on track if scenario elements were missed. These tended to be ad-hoc and break some of the fluency of work. More dedicated efforts such as the use of adaptive white cells were seen employed in the military and EOC cases. However, these also share the same potential pitfall of adapting events to simply recover from immediate challenges while losing the big picture or potentially jeopardizing later planned for events. As was discussed, the more ad-hoc these adaptation are, the more difficult it is to trace to objectives both upstream to learning and downstream back to planned intent. When a failure occurs or work is undertaken that is not expected, the run-time event can shrink and result in lots of people standing around doing nothing. One has to be able to re-engage them not just for the pacing, but so that the participants can resume agent roles and interactions in the simulated field of practice.

- *Learning lab functions are reminders to plan ahead and cater to multiple perspectives and goals across levels of learning.*

Stakeholders and exercise designers need to be able to inject new ways to get information about critical objectives that are planned. When a probe event does not work, designers and observers need ways to adapt scenario events and still learn about the intent behind the probe event. When a pacer or probe event is lost, or a new event occurs, there must be ways to reactivate and get things moving. The learning lab functions are specific ways to think about planning ahead by incorporating such events in the planning and run-time phases. Planned reflective forums and rally points have a naturalness that can be synthesized into the course of a run-time if designers need to cope and recover from such kinds of disruptions. Each learning lab function serves
as information to the multiple roles and perspectives of stakeholders and observers. A reflective forum for example, is a tool that not only externalizes cognition for a CSE perspective, but it also serves as a landmark in the exercise from which to analyze not just the specific case but can also provide insight to general issue to other problem holders as well. To explain that further, in urban firefighting for example, the general case is often seen is that of anomaly response. Within this general case, one can immediately see traces and examples of hypothesis generation situations (such as in problem recognition situations). Such analysis is not just about the diagnosis in that particular domain, but rather information from analysis can be abstracted to reflect on hypothesis generation in any situaiton.

When looking at training and learning, training is local and based only on the events and actions that actually happened. Training and instructional objectives are often seen as drilling specific skills for specific cases. Successful learning however goes beyond just having the people, the case, and the specifics all in one place. Learning can be about training, as was seen throughout the interview chapter where CSE efforts were often piggybacked on training, but as a learning lab, there are multiple other purposes that can be served. Given the logistics efforts put into pulling these events off, it means there is a learning opportunity that can be taken advantage of. The criterion for success is whether our planning and adaptation efforts widen the learning cycles between different groups. The learning lab functions, in the form of rally points, pacers, probes, and reflective forums, are not just elements for occasional use, but are critical functional components to support these multiple perspectives and varied learning goals.
Coordinated observational agility and protocol analysis ‘on the fly’, helps abstract conceptual findings and re-direct attention to unplanned events.

A profound finding throughout all the exercises and staged world studies discussed, is the power of observation. As was explored in the Huachuca case, the ability to direct observational resources to get the right data to analyze is imperative. What is novel though, is that while good pre-planning will help prepare observer knowledge of what the interesting data will be and help set up observation in advance, what is imperative is the observational ability to re-focus and re-orient to surprise and new potentially valuable events. This ability of going into an observation, focusing on one set of events or probes and then recognizing the need to switch to another highlights this need to reshuffle observational researcher perspectives. In the smaller bounded scale cases such as the nuclear simulations discussed in Chapter 3, it was possible to record everything from a single team in an hour or two. Given the scale and scope of large-scale exercises, this is not as tenable and its the coordinated flexibility that is now crucial.

In both the Huachcua case, the Kings Mall case, and the Dayton EOC case, they emphasize the need for an observational roadmap, the need to plan the probes, and the need to be able to say what concepts are really being pursued because it allows one to focus their observation and come out with a reasonable analysis that could be checked. This ties in with the need for contingency planning. The events and interactions that create contingencies- such as missed or ceiling/floored probes, may cause something unexpected to become a new probe. Observation needs to be able to recognize that this has happened and adapt the observational scheme to get those data records. What became critical to accomplish this, as seen in the cases from
Chapter 2, was the ability to use synchronous and asynchronous communications to help perform protocol analysis on the fly to help direct observational resources. Protocol analysis is needed to manage this because observers need to be able to recognize these situations as opportunities. As mentioned in regard to the narrow experimentalist approach discussed in Chapter 2, simply defining coding in advance, doing the coding on an event, and then looking at inter-rater consistency is not a resilient observational approach.

**Intersection of Contributions**

These contributions are mutually reinforcing and the planning tool is designed to support their synthesis. At the intersection of these contributions is the imperative critical support provided by a centralized observation center that pulls together an account of how the learning lab is unfolding which can then be used to assure the run-time exercise is doing the best job it can given the way events and adaptations are unfolding relative to meeting the stakeholder objectives. As well, they become critical in developing purposes behind anticipating what is going to happen at run-time, and tracking out how that gets disconnected. By sharing tentative observations across the distributed observation team, the observation command center is responsible for synthesizing an account of how a run-time exercise is coming together, how practitioner teams are doing relative to difficulties encountered, and facilitating the switching between the language of the field of practice and the conceptual language in any field of practice. As the central observation coordinates and disseminates a real-time account of how the run-time is unfolding and where difficulties are occurring, this can be related to multiple objectives from instructional local objectives to traces of higher
level CSE patterns. Central observation needs to be able to tell when instructional goals or multiple other stakeholders are not being met in order to re-plan.

Such dynamic re-planning cannot be afforded by mere prescriptive checklists and normative procedures prior to the event, but rather the centralized observation command must be able to coordinate with distributed observers coupled with a real-time display to visualize the longshot of the exercise roadmap and quickly see what is potentially being lost as well as the intent behind individual parts of the staging process. The initial Learning Lab Designer prototype is one such design approach to capture such a longshot. By putting the emerging run-time observation into this context, new events and adaptions will reveal what is being gained as well as help designers and observers see when they might need to backtrack and explain why events were there in the first place. With this longshot, observer coordination and exercise re-planning decision making can be facilitated. If something new comes up or a surprise breakdown occurs, by seeing the whole picture in context, planners and observer controllers can make better more planful decisions (such as by turning a rally point into an opportune reflective forum) to re-steer the run-time in a face valid way that preserves learning goals and intent.

As was mentioned with regard to contingency planning, this often has to be performed because the exercise may start to go off-track or ‘lose steam’. By coordinating around analytical results to manage the run-time exercise, the learning lab functions coupled with the longshot provide a planful support technique to fulfill this need to regenerate or re-engage participants within the exercise. As with re-planning, when there’s an impasse to a plan in progress, one needs to be able to see how intent is brought in [Shattuck & Woods 2000]. This can be relevant to a wide
range of applications not only as a research result that can be looked at in many
cognitive systems, but also as a guide to specific situations of how this world can
see how communication of intent can lead to breakdowns in plans. When something
doesn’t work, the distributed real-time protocol analysis results of conveying what is
interesting observationally has a role in the cognitive model of what makes challenges
and work contexts difficult. By sharing tentative observations and focusing on the
collaboration, given this is distributed observation, instead of one person going over the
data separately, multiple people are sharing, collaborating, and re-tuning. This sharing
becomes critical because the groups responsible for observing and conducting the
exercise need the partial protocol analysis to match the contingency decision making
in the run-time exercise, re-direct observational resources to get good data capture on
interesting parts, and to enhance the efficiency and effectiveness of the analysis after
the events are over, because there are always limited analysis resources given the scale
and scope of the data one gets in these types of exercises. As we abstract out and
synthesize findings from an exercise and iterate across the learning cycle to sharable
and generalizable learning, these learning lab functions and contributions are set up
to encourage and facilitate that process.

5.1.1 Learning Lab Abstraction

Throughout the course of this dissertation, across personal research involvement
investigating staged world design with expert researchers and practitioners, and through
identifying challenging support needs, a number of insights have been made to help
design effective staged world studies for learning. The learning lab environment is an
attempt to provide guided structure and reflection on the many CSE processes behind
getting valuable observations and lessons from field research, and placing the scenario
design of staged worlds at the nexus of this process. The series of learning lab figures
thus far have shown how to develop an exercise, but pivoting that figure, we can assess
a new level of abstraction from the learning lab design. In the spirit of Rasmussen’s
original abstraction hierarchy (Rasmussen, Pejtersen, & Goodstein [1994]), the learning
lab framework in Figure 5.1 can be seen from a different perspective representing
an overall basic functional goal decomposition. Several layers of representation are
necessary to relate the two extremes in the abstraction—(1) generalizable and sharable
learning derived from (2) field exercise scenarios. As one shifts up and down these
levels of abstraction, the concepts used to represent the functional structure changes.
Scenarios and story-lines become the transformational links between as well as the
pivotal exercise run-time itself at the bottom of the exercise.

The ultimate goals of the learning laboratory are at the top, and this is the sharing
and dissemination of generalizable knowledge to a research base. Incidentally, it is
this same research base that forms the purposes and functions that start the planning
for an exercise. Unlike a traditional abstraction hierarchy where details of information
are aggregated up across each layer, in this simplified take on Rasmussen (1994) the
top level principles in this case, ‘insight and discovery’ are at the same level as the
purposes and functions that go into the planning. In this sense, the exercise run-time
can be seen as being derived from the purposes and functions behind its design, but
these are equally influenced by the learning that is occurring in operations.

At the bottom, the exercise run-time is the physical form of a scenario used for an
event. Practitioners, stakeholders, and observers all took part in activities in this event,
and each of those perspectives represent story-lines of practice. The activities this
Figure 5.1: Looking at a cross-section of the learning laboratory upon exercise completion, one can now visually trace and see alignment between stakeholder objectives and actual outcome.

practice was based on comes from specific learning laboratory functions and a script which represented hypotheses about scenario elements. Now the representation of those activities, via process traces or animock blocking diagrams capture the local knowledge used by practitioners in situ, leading into the abstract processes and patterns of work. Learning lab functions, such as probes should be able to tap into these hypothesized functions in the exercise run-time. Insight & Discovery is the top level goal which now has tracability from this high-level goal all the way back to what was seen in the exercise, which can be traced even further back to how, what, and why it was
there in the first place with the purposes and functions. In this sense, the exercise and the scenario are not just the system and the human field of practice, but rather the exercise design as a learning lab is the physical artifact, that is used to assess JCS performance. It is this new learning lab artifact and the balance across dimensions that tells new stories about future designs which will influence the creation of new scenarios and exercises.

5.2 Final Remarks and Conclusion

This dissertation has illustrated the complexities of designing staged world studies given the new challenges introduced by various dimensions of scaling. The learning lab framework takes a CSE approach to balancing practice-centered research with design. Complementing earlier CSE tools for collaborative envisioning such as the animock (Roesler et al., 2001) and the topic landscape (Schoenwald et al., 2005), the learning laboratory concept is a robust way to approach the design of staged world planning and observation.

The data abstracted from a variety of domains including crisis management observations and from a series of interviews with key researchers suggests there are a number of difficulties throughout the planning, staging, and evaluation of large scale exercises. The learning lab approach examines many of these challenges from multiple perspectives to help future designers and researchers effectively capture and explore learning opportunities. In reviewing CSE’s own history in how the field has used exercises as learning labs, one finds that these studies have always had a number of scaling components and this work aims to unearth some of the new collaboration and coordination challenges evident as we move into new domains and new scales.
A planning tool design seed was developed to support resilient processes for conducting staged world exercises. Planning needs to be resilient because when one scales up, given the use of distributed groups, spatially separated agents, and multi-hour and multi-day work environments, unanticipated disruptions and changes will make sure events do not go as planned. The planning tool is not simply about making analysis more efficient, but rather staging the conditions for learning. The previous generation of planning tools were largely designed to overcome technical logistics to make collection easier. They were focused on trying to help with local learning and then it went on to move to more generalizable forms so one could talk in more abstract ways. However, these approaches never really addressed the planning side. These tend to assume researchers already have everything in control in the staging, and the interactions across stakeholders and multiple purposes have already been established. The Learning Laboratory Designer tool was developed to be a planning tool from the start. The issue is not so much on the analysis side, but rather it supports how to stage such exercises so that researchers will be able to learn and abstract from these field observations. Because there is a real run-time exercise, real events are supposed to be happening to people who are supposed to be running in the event—because of this it’s easy to get lost in the details. In order to provide the longshot that helps stakeholders and planners step back and take in a big picture, the tool explicitly maps elements back to intent to help those individuals make better decisions.

Looking back, it becomes interesting that studying disaster response by using ongoing training exercises and by creating new exercises, the process is a lot like handling a disaster itself. The run-time exercise will not unfold as planned and specified, communications and equipment will fail, and the situation will not allow
individuals to learn in the way that they thought they were going to learn. The tools that I have have used and claim would be helpful in the dynamic and distributed world of emergency management, turn out to also be releveant to organizing, planning, and staging disaster exercises. This highlights the need for new tools and artifacts that can effectively supporting pulling observations together and sharing them such that everyone involved has a common grounding. The approaches we have successfully used and demonstrated in this dissertation around developing distributed collaboration and coordination systems need to be applied to the planning and execution of these large scale exercises. The value of this is that it dramatically increases the potential for learning related back to the main empirical finding from this work- that relative to the scale of resources that have to be committed to execute these types of exercises, the amount of learning yield so far has been remarkably low. For the future planners, designers, developers, and researchers building the next generation of exercises and support systems in mission critical domains, they must be able to effectively adapt to and overcome such challenges as those suggested in this dissertation in to order to support resilience and develop a deeper understanding of cognition at work in complex domains.
LIST OF REFERENCES


