THE CYBER DEFENSE REVIEW

The Concept of a "Campaign of Experimentation" for Cyber Operations

Author(s): Robert R. Hoffman

Source: *The Cyber Defense Review*, Vol. 4, No. 1 (SPRING 2019), pp. 75-84

Published by: Army Cyber Institute

Stable URL: https://www.jstor.org/stable/10.2307/26623068

The Concept of a "Campaign of Experimentation" for Cyber Operations

Dr. Robert R. Hoffman

ABSTRACT

A Campaign of Experimentation is necessary for the United States to achieve a robust capability in cyber defensive and offensive operations, that is effectively and efficiently integrated with operations in cyber-kinetic domains. The article describes challenges for such a Campaign, regarding experimental design, logistics, measurement, and methodology.

The campaign concept

In a report titled "Code of Best Practice: Experimentation," David Alberts and Richard Hayes [1] asserted:

> Experimentation is the lynch pin in the DoD's strategy for transformation. Without a properly focused, well-balanced, rigorously designed, and expertly conducted program of experimentation, the DoD will not be able to take full advantage of the opportunities that Information Age concepts and technologies offer.

Alberts and Hayes continue to explain why the DoD needs to conduct "Campaigns of Experimentation." First, no single experiment improves knowledge enough to support a major goal such as transformation. Individual experiments can only look at a limited number of variables and contexts, and therefore must be integrated with other experiments to ensure that limiting conditions are properly understood. Series of experiments are needed to differentiate between competing hypotheses to yield actionable knowledge. Second, individual experiments within a series are likely to generate some unexpected findings that are both important and interesting. Experimentation campaigns provide the opportunity to explore those novel insights and findings, as well as their implications. These ideas are all fundamental to the methodology that has been established in the field of experimental psychology. [2]

Robert R. Hoffman received his Ph.D. in experimental psychology from the University of Cincinnati. He is a Senior Member of the Institute of Electrical and Electronics Engineers (IEEE), a Fellow of the Association for Psychological Science, a Fellow of the Human Factors and Ergonomics Society, a Senior Member of the Association for the Advancement of Artificial Intelligence, and a Fulbright Scholar. He has been recognized internationally for his research on the psychology of expertise, the methodology of cognitive task analysis, and the issues for the design of complex cognitive work systems, including cyber work systems.

We are now in what Alberts and Hayes referred to as the "Information Age Transformation". [3] While the scientific rationale for a Campaign of Experimentation is based on the above considerations, the practical rationale is equally significant. The current world situation is one in which adversarial relations and conflicts are characterized by extreme levels of uncertainty, complexity, fast pace, and dynamics. The delivery of a technology or weapon system is not the end of a procurement. It is the beginning of a phase in which operations and experimentation must be tightly coupled. What this means is that the traditional separation of experimentation and operations must not just be blurred but dissolved. As ever-more complex automation is injected into the workplace, the work must be continuously observable.

What makes cyber operations unique

The domain of cyber operations is unique in several respects, which further justifies the application of the Campaign concept. Though network operations is not a new type of work, the work of U.S. Cyber Command (USCYBERCOM) Cyber Protection Teams (CPTs) is relatively new. While there are many experts who have considerable experience in network operations, many of them work in the private sector. For CPT certification and performance evaluation there remains a gap in our ability to appropriately describe the work in terms of proficiency scaling and learning curves. It is not enough to say that an individual has specific qualifications as evaluated by a checklist method. One needs a full and rich description of what it means for an individual to be an apprentice, journeyman, expert or master. We know this to be true for all other complex sociotechnical domains. [4]

Cognitive work in the cyber domain is a moving target as it involves an adaptive and deceptive adversary and a rapid pace of technological change. The pace of change in the work and the technology far outstrips the speed at which standard controlled experimentation can be conducted. As both cyber work and cyber tools continue to evolve and cyber Concepts of Operations (CONOPS) continue to adapt, there is the need to understand the issues and provide recommendations to ensure an effective cyber force. Mission types will change as threats and adversaries themselves change and adapt. Research must be ongoing.

Cognitive work is messy. Numerous uncontrollable variables come into play and can influence logistical and operational activities. Were experimentation to be conducted in the traditional manner of isolation and control of variables, the research would not represent the actual work ecology. Tasks that are tightly bound by procedure when conducted in the laboratory might permit careful measurement, but can also distance the task process from real world variables. Thus, research is needed that combines both laboratory and field experimentation.

There are more variables that play a crucial role than can be controlled and manipulated in any single experiment: the experience level of the cyber workers who are research participants, the technologies utilized, the different sorts of missions, and the various logistical demands that must be met. Research designs can adopt any of several options, ranging from single, simple experiments that evaluate baseline performance, to larger, more complex designs that involve the manipulation of more than one variable. There must be an on-going process of developing useful experimental designs and mapping them on to the immediate needs that emerge.

There is no clear or straight path from high-level concepts such as “efficiency” and “quality” to operationally defined measures that are useful in experimentation and evaluation. Cyber operations involve multiple sub-tasks. The tasks and sub-tasks are not strictly linear or stepwise but are often conducted in parallel. [5] These and other features of cognitive work mean that experimentation is necessary to develop and refine appropriate measures and metrics.

Concepts for experimentation on cyber work processes and tools challenge our fundamental notions of statistical testing and analysis. A primary reason is sample size due to resource limitations. Suppose, for example, that one has a new software tool suite to evaluate. The evaluation must involve multiple cyber operators attempting to learn and use the new software tools, but multiple cyber operators are often not available. And when they are, they must be selected for having similar levels of experience, which means that experimental designs based on traditional parametric statistical significance testing can be insufficient. Therefore, methods of order statistics and concepts of practical significance must be considered. It should be noted that a Campaign of Experimentation represents a unique and important opportunity to advance our scientific methodology for statistical analysis of studies having small sample size, and for the large-scale experimentation that is resource

constrained. In addition to mandating advances on the concept of practical significance, it is necessary to make advances on the estimation of effect sizes given small sample sizes. [6][7]

The above considerations all mandate a Campaign of Experimentation as an on-going process. The Campaign would be conducted not only to address the above needs but to also recognize a fundamental fact of scientific experimentation: that the purpose of experimentation is to continually improve and refine the experimental and measurement methods.

A Campaign of Experimentation is necessary to inform cyber CONOPs. Research evaluates the technologies and software systems for their understandability, usefulness, and usability. The performance of cyber operators must be empirically observed and evaluated to ensure that the work is effective and is of the highest quality. Research shapes our understanding of proficiency levels for selection and training.

Moving from the campaign concept to a cyber-specific methodology

Alberts and Hayes [3] presented some "barriers to transformational campaigns." For instance, they cautioned against the imposition of unrealistic schedules on experimentation, the failure to utilize an extensive and rich set of realistic scenarios, and the failure to adequately fund the experimentation. While expressing such important cautionary tales, the work of Alberts and Hayes did not delve deeply into the procedural and methodological details involved in experiments of the sort being envisioned, specifically experimentation on Cyber Operations.

However, results from recent research activities at the Cyber Immersion Laboratory of USCY-BERCOM have illuminated several vital principles that take the broad Alberts-Hayes concepts and apply them specifically to Cyber Operations. The NetMap activity [8] and the Deployable Mission Support System (DMSS) activity [9] engaged CPTs in processes of network mapping and vulnerability analysis. The purpose of these activities was to observe and evaluate the performance and workflows of CPTs, observe and evaluate the usability and usefulness of the available software support systems and tools, and initiate a process of capturing the knowledge and reasoning strategies of the most experienced CPT members. These activities required the establishment of a virtual cyber environment, the scripting of various scenarios, the coordination of multiple CPTs, and other logistical elements required for large-scale experimentation.

The process of designing, implementing and conducting these activities revealed many challenges. For example, it was determined that each CPT member would have to complete a demographic survey, complete various checklists as they accomplished sub-task goals, complete a post-event questionnaire, among other tasks that are not a part of regular CPT activities. Once the requisite materials were fleshed out and used, it became clear that the participants were in some sense being over-burdened. Clearly, the experimental context should not demotivate the participants. Several additional challenges emerged from these projects.

Experiment design challenges

Experiments require that some variables are controlled while some are manipulated. The manipulated variables are the ones whose causal impact is of immediate interest. The controlled variables are the ones that are known or believed to have an impact, but that must be held constant for the assessment of the manipulated variables. For instance, one might want to conduct an evaluation of a software tool but hold participant experience level constant by involving only the highly experienced CPTs (a control variable). One might want to have CPTs work on more than one type of attack (a manipulated variable) to evaluate task difficulty.

There are more important variables that can be manipulated and controlled than can be logistically incorporated. Take the example of task difficulty. A CPT conducts a task (e.g. vulnerability analysis) using software Tool A and then repeats the task using Tool B. But in using tool A the first time, the CPT will have become familiar with the network under study, perhaps making it only seem as if they perform better on the second task. This means one needs a counterbalanced order, in which one CPT uses Tool A first and the other CPT uses Tool B first. The alternative is to build more than one test network. Then, there is the matter of CPT experience. Do we want to make decisions about tool usability based on the performance of trainee CPTs or based on the performance of experienced CPTs? However, one approaches the design challenge, the experiment design can quickly become complicated.

Another design challenge is that the findings from a highly controlled environment might not apply in messy real-world instances where the work involves many uncontrolled and uncontrollable variables. If one wants to know about such things as CPT performance or tool usability, then those things must be evaluated in ecologically valid and varied conditions rather than in tightly controlled environments in which key variables get frozen out. Experiments must let the nasty variability of the world enter the picture. This runs counter to the traditional paradigm of laboratory experimentation. It is therefore crucial for a Campaign to involve specialists who have had experience in laboratory experimentation, and who can take point on matters of experimental design and measurement.

Logistic challenges

The challenges of experimental design mentioned above spill over to logistics. A counterbalanced design involving, for example, high and low experience CPTs, multiple software tools, and the need for multiple test networks, etc., means mustering human and machine resources that can be hard to come by. That nasty variability of the real world can entrain considerable logistical problems. For example, even simple things (such as failure of a disc to initialize) can completely shut down a large-scale experiment and send 50 people home. Just as unexpected things happen in the "real world," unexpected things happen in the context of large-scale experiments. This is something that the researchers must navigate.

Measurement challenges

There is a tendency for researchers to seek easy, automation-based methods for collecting data. In the case of cyberwork, for instance, this might involve examining logs of operator actions. But log data do not inform you about what the operator was thinking, anticipating, or worried about. Logs would tell you something about what they were doing, but not why they were doing it. Another measure that is often mentioned is eye movements. Eye movements may tell you what an operator is looking at, but they do not always tell you what the operator is thinking or is worried about.

There is a distinction between objective and subjective data, coupled with the mistaken belief that subjective data do not make for genuine science. It has been argued in the philosophy of science for decades that the distinction between objective and subjective data is mythical; all measures have both subjective and objective aspects to them. [10] Cyberwork is deeply and necessarily cognitive. The analysis of CPT members' reasoning and knowledge is central to the development of an effective workforce and can only be evaluated if the researches somehow "get inside the heads" of the CPT members, primarily by asking questions in structured cognitive interviews. [11] The most important data always come from the participants' answers to probing questions. What are you thinking? What are you anticipating? What are you worried about? What is your machine doing?

The drive to find useful metrics brings in another measurement challenge. Certainly, meaningful measures are needed, including measures that can be taken automatically, but this is not the same as metrics. A metric is a decision point, a value on some measurement scale that informs decision making. Is a score of 70% correct indicative of good performance, or poor performance? Well, it depends on the task. Metrics do not derive directly, easily, or automatically from measures. Theories provide measurable concepts, and measures are recipes for taking measurements, but metrics come from policy. [12]

The drive to automate measurement, and the belief that automated measures are objective, combined with the belief that all scientific and policy answers can be found if only if one has good metrics are all beliefs that blind researchers to the fact that research is difficult.

Methodological challenges

It is important to keep in mind one of the purposes of experimentation and measurement is to continuously adapt and improve the experimental and measurement methodologies, especially in the Campaign context where events provide opportunities that could be easily missed. For instance, there may be a lull in the cyberwork activity (for any of a variety of reasons). From a research perspective, lulls are an opportunity to conduct cognitive interviews with the CPT members to assess such things as their training and development of expertise, to elicit information about their reasoning strategies, and the experience that enabled them to achieve expertise. They can be asked about how they learn the differences

between their actual work process and doctrine (i.e. lessons learned and best practices), the tool functionalities and capabilities that they need or desire. [13]

Based on experience in the NetMap and DMSS projects, recommendations can be offered concerning methodology. First, it is recommended that a Dry Run study be conducted before the actual experiment activity. In a Dry Run, the researchers themselves serve as cyber operators and attempt to conduct the tasks, using a highly scripted workflow. The purpose is to evaluate the planned experiment procedure and familiarize the researchers with the workflow.

The second activity is a Pilot Study. A select, highly-experienced CPT conducts the experiment procedure while researchers observe and present probe questions. The purpose is to evaluate the planned experiment procedure and familiarize the researchers with the workflow, but also to forge an all-important performance baseline.

Third, experiments need to have a Conductor, a selected researcher who issues directions to observers and CPT operators, starting in the scripted dry run and continuing in the pilot study. The Conductor keeps things coordinated and gains an appreciation of where the planned experiment procedure falls on the continua of complexity and ecological validity. Experiments of the sort that have been referenced in this article involve upwards of six researcher/observers, five CPTs, and additional support and technical staff.

Fourth, there must be a deliberate effort to build a useful baseline, which would start with the Pilot Study and continue through to the Baseline Study, in which a select, highly experienced CPT engages in the experimental procedure and tasks without any direction, scripting, or interference from the researchers. The purpose will be to evaluate the ecological validity of the scripted workflow and procedure and refine the performance baseline and other measures.

CONCLUSION

The need to develop a U.S. Government Cyber CONOPs and capability based on the Alberts-Hayes concept of a Campaign of Experimentation is apparent. It is in some respects being implemented in current studies at facilities such as the Cyber Immersion Lab of USCYBERCOM and the U.S. Army's Cyber Human Integrated Modeling and Experimentation Range. To some extent, the Campaign concept is being partially implemented in various cyber events, exercises, and competitions. The purpose of this article is to motivate a programmatic process for fleshing out and fully implementing the Campaign concept with specific reference to the unique needs and challenges of cyber work. A broader implication of the challenges presented here is that the full implementation of the Campaign would require the coordinated integration of resources and activities across several branches of government, including but not limited to the Department of Defense.

Another challenge that should be noted involves both logistics and experimentation. Since a Campaign can span years, and the individual experiments can span months in planning and implementation, it is crucial for there to be continuity of the Campaign leadership. A stable vision accompanied by a deep understanding of the Campaign and its individual projects and experiments will be necessary for Campaign success.

DISCLAIMER

The views and conclusions contained herein are those of the author and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government.

NOTES

1. D.S. Alberts and R.E. Hayes, (2002), "Code of Best Practice: Experimentation." Command and Control Research Program, Department of Defense, Washington DC, xi.
2. B.J. Underwood, (1966), Experimental psychology. New York: Appleton-Century-Crofts.
3. D.S. Alberts and R.E. Hayes, (2006), "Code of Best Practice: Pathways to Innovation and Transformation." Command and Control Research Program, Department of Defense, Washington D.C., 136-139.
4. R.R. Hoffman, P. Ward, L. DiBello, P.J. Feltovich, S.M. Fiore, and D. Andrews, (2014), Accelerated Expertise: Training for High Proficiency in a Complex World. Boca Raton, FL: Taylor and Francis/CRC Press.
5. S. Trent, R.R. Hoffman, D. Merritt, and S. Smith, (in press), Modeling the cognitive work of cyber protection teams. The Cyber Defense Review.
6. R.R. Hoffman, P.A. Hancock, and J.M. Bradshaw, (2010, November/December), Metrics, metrics, metrics, Part 2: Universal Metrics? IEEE Intelligent Systems, 93-97.
7. R.R. Hoffman, M. Marx, R. Amin, and P. McDermott, (2010), Measurement for evaluating the learnability and resilience of methods of cognitive work, Theoretical Issues in Ergonomic Science. Published online at iFirst, DOI: 10.1080/14639220903386757.
8. S. Trent, R.R. Hoffman, and S. Lathrop (May 2016), Applied Research in Support of Cyberspace Operations: Difficult, but Critical, The Cyber Defense Review, https://cyberdefensereview.army.mil/CDR-Content/Articles/Article-View/Article/1136076/applied-research-in-support-of-cyberspace-operations-difficult-but-critical/.
9. S. Barna, (with 9 others) (2017), "Deployable Mission Support System Assessment." AOS Report No. AOS-17-0524, Applied Physics Laboratory, Johns Hopkins University, Laurel, MD.
10. F.A. Muckler, (1992), Selecting performance measures: "Objective" versus "subjective" measurement. Human Factors, 34, 441-455.
11. B. Crandall, G. Klein, and R.R. Hoffman, (2006), Working Minds: A Practitioner's Guide to Cognitive Task Analysis. Cambridge, MA: MIT Press.
12. R.R. Hoffman, (2010), Theory → Concepts Measures, but Policies → Metrics. In E. Patterson and J. Miller (Eds.), Macrocognition metrics and scenarios: Design and evaluation for real-world teams, London: Ashgate, 3-10.
13. R.R. Hoffman and M.J. McCloskey, (2013, July/August), Envisioning Desirements, IEEE Intelligent Systems, 82-89.