INSTITUTE FOR DEFENSE ANALYSES

ITEA Editorial – A Groundswell for Test and Evaluation

Laura J. Freeman

October 2018

Approved for public release.

IDA Non-Standard Document
NS D-10324

Log: H 2018-000457

INSTITUTE FOR DEFENSE ANALYSES
4850 Mark Center Drive
Alexandria, Virginia 22311-1882

The Institute for Defense Analyses is a non-profit corporation that operates three federally funded research and development centers to provide objective analyses of national security issues, particularly those requiring scientific and technical expertise, and conduct related research on other national challenges.

About This Publication
This work was conducted by the Institute for Defense Analyses (IDA) under contract HQ0034-14-D-0001, Task BL-9-4555, "Reform Management Group Technical Advising," for the Deputy Chief Management Officer. The views, opinions, and findings should not be construed as representing the official position of either the Department of Defense or the sponsoring organization.

Acknowledgments
The IDA Technical Review Committee was chaired by Mr. Robert R. Soule and consisted of Jane K. Yevgeniya and Matthew R. Avery from the Operational Evaluation Division.

For more information:
Laura J. Freeman, Project Leader
lfreeman@ida.org • (703) 845-2084

Robert R. Soule, Director, Operational Evaluation Division
rsoule@ida.org • (703) 845-2482

INSTITUTE FOR DEFENSE ANALYSES

IDA Non-Standard Document NS D-10324

ITEA Editorial – A Groundswell for Test and Evaluation

Laura J. Freeman

ITEA Editorial - A Groundswell for Test and Evaluation

The fundamental purpose of test and evaluation (T&E) in the Department of Defense (DoD) is to provide knowledge to answer critical questions that help decision makers manage the risk involved in developing, producing, operating, and sustaining systems and capabilities. At its core, T&E takes data and translates it into information for decision makers. Subject matter expertise of the platform and operational mission have always been critical components of developing defensible test and evaluation strategies. Recent innovations in data science have improved our ability to collect, store, manage, transfer, process and visualize data. Additionally, advances in statistics and uncertainty quantification are revolutionizing how we think about predictions from all types of data. The ability to integrate system and scientific knowledge, coupled with advances in data science and statistics, will enable us to better target testing, make efficient use of resources, quantify risk, and lead to well informed decisions.

Change is in the Air

Over the past year, through my roles as Senior Technical Advisor to DOT&E and now as analytical support to the Test Reform Management Group, I have interacted with senior leaders from across the Office of the Secretary of Defense (OSD) and the Services. There is enthusiasm for updating T&E processes and methods to reflect recent advances in computer models, big data, analytics, and statistical methods for using all available information.

Some recent activities in OSD supporting change:

- Dr. Michael Griffin, Undersecretary for Research and Engineering, signed the Digital Engineering Strategy.[1] Engagement between T&E and digital engineering focuses on how T&E provides input to the authoritative truth for system performance data.
- Department of Defense Chief Management Officer (CMO) created the Test Reform Management Group (TRMG). Focus areas related to this article include "Shift Left" T&E, integrated testing, mission-based design of experiments (DOE), and planning for future technologies.
- Mr. Robert Behler, Director Operational Test and Evaluation (DOT&E), in his first Annual Report[2] expressed a desire to expand integrated testing and use more modeling and simulation for evaluations. Director Behler noted that while the current efforts to integrate testing were an important first step, we should go further and incorporate relevant aspects of the mission earlier in developmental testing. He pledged to use all credible data in DOT&E's evaluations.
- The Test Resource Management Center (TRMC) highlighted Big Data Knowledge Management in its strategic plan.

While these initiatives are independent activities, they have common goals. Themes I have noted are 1) the desire to use new technologies to generate more information early in the acquisition lifecycle (e.g., computer models and simulations), 2) the desire to ensure we collect the best

[1] https://www.acq.osd.mil/se/initiatives/init_de.html

[2] http://www.dote.osd.mil/pub/reports/FY2017/

possible data in T&E from the inception of testing, and 3) the desire to use all credible information for decision making.

Data Science Advances

In a recent address at the Fall Technical Conference,[3] speaker Dr. L. Allison Jones-Farmer provided an insightful review of the history of data science and highlighted key characteristics of successful data science companies. The DoD T&E community should take note that companies leading the field of data science have the following characteristics:

1. The ability to define infrastructure and analysis tools
2. Ownership of the data
3. Data-drive decision-making processes

As a tester, these points resonated with me. TRMC has a strategic plan for how we define and build our own infrastructure to support T&E needs, considerations include data access, storage, and built-in analysis tools. Data ownership presents unique challenges in the DoD T&E community. For example, if the program owns test data, how do we facilitate data use across programs or Services? Additionally, classification and special access considerations complicate use of data. We need to strategically plan for data ownership and use to address these challenges. Finally, as testers, it is our job to communicate our findings to decision makers in a way that they can use it to inform their decisions. We must develop visualizations that distill the vast amount of information available for our decision makers. We should put tools and practices in place to define how we collect, store, and process data.

There has been substantial progress in the development of tools and methods for data capture, storage, transmission, curation, and visualization. In their 2014 review paper,[4] Chen and Zhang review over 200 references and highlight tools, methods, and challenges for big data in each of these areas. One principle they highlight is that for big data, we need to bring the analysis to the data instead of the standard approach to distribute the data to stakeholders. This will be a paradigm shift for the DoD.

Industry has also found that data quality should not be ignored. Google recently released a new data set search tool and guidance to data set providers to improve search outcomes.[5] Datasets are "easier to find when you provide supporting information such as their name, description, creator and distribution formats as structured data," according to Google. While this is certainly a best practice that the DoD should follow, we must also note that the information extracted from data is only as good as the context provided. Dr. Roger Hoerl and Dr. Ron Snee have defined a

---

[3] http://www.falltechnicalconference.org/program/

[4] Chen, C. P., & Zhang, C. Y. (2014). Data-intensive applications, challenges, techniques and technologies: A survey on Big Data. Information Sciences, 275, 314-347.

[5] https://www.blog.google/products/search/making-it-easier-discover-datasets/

structure for understanding the data pedigree.[6] They define the following as essential for understanding the pedigree of data:

- What the data represent; that is, a basic explanation of the underlying subject matter knowledge of the phenomenon being measured, including units of measurement.
- Description of the process that produced the data.
- Description of how the "samples" were obtained from this process that were subsequently measured.
- The specific measurement process used to assign numbers or attributes to the "samples."
- The existence (or lack) of recent analyses of the said measurement system, such as gage repeatability and reproducibility (R&R) studies, calibration studies, and so on.
- The history of the data, documenting the chain of custody - who has had access to the original data, and what, if any, changes or deletions have been made - and access to a copy of the original data that can be verified.

The history of your data is essential in order to reproduce analyses and reuse data for years to come. I am sure we have all shared similar experiences: spreadsheets with data are passed around, errors are introduced, and hours are lost trying to get to the authoritative source of data. With big data, the problem grows exponentially. The DoD needs to adopt a data pedigree standard moving forward.

Statistics Advances

As a card-carrying statistician, I could talk about useful methods for pages on end. However, I will limit this section to statistical methods that I believe will ultimately have the largest impact on the T&E profession. A theme for these methods is that they capitalize on subject matter expertise, but also allow flexibility for learning as we acquire test data. These are qualities that the best methods employed in T&E will share moving forward.

Bayesian methods – In their Quality Engineering paper on Bayesian reliability, Dr. Alyson Wilson and Dr. Kassandra Fronczyk highlight that, "One of the most powerful features of Bayesian analyses is the ability to combine multiple sources of information in a principled way to perform inference."[7] While the paper focuses on reliability, the concepts apply to all types of test data. Moreover, Bayesian analyses provide the flexibility to incorporate scientific knowledge, operational mission knowledge, and other subject matter expert information using a formal methodology. As the DoD T&E community seeks to improve our process, use all available information, and even combine information from computer models and simulations with live test data, Bayesian methods must become part of our regular toolbox.

---

[6] https://www.researchgate.net/publication/290547300_Inquiry_on_pedigree

[7] https://www.tandfonline.com/doi/abs/10.1080/08982112.2016.1211889

Design and Analysis of Computer Experiments - Advances in computing, scientific understanding, and mathematical modeling, have made it possible to use computer simulations to augment live testing.[8] The field of design and analysis of computer experiments provides new methods for spanning complex inputs spaces for simulations, and provides the analysis tools to create surrogate models (often referred to as emulators) to represent the outputs of those simulations. These design and analysis methods give analysts tools to both design tests to focus on critical information and identify edges of the operational mission space, two goals that DOT&E has emphasized since 2002.[9]

Uncertainty Quantification - The National Research Council[10] defined uncertainty quantification as "the process of quantifying uncertainties associated with model calibrations of true, physical quantities of interest, with the goals of accounting for all sources of uncertainty and quantifying the contributions of specific sources to the overall uncertainty." They developed a new acronym for Verification, Validation and Uncertainty Quantification, (VVUQ), which emphasizes uncertainty quantification as a critical aspect of the V&V process. There are many possible sources of uncertainty that the T&E community must consider, especially when combing results from computer models with live test data:

- Computer model input uncertainty
- Computer model parameter uncertainty
- Model inadequacy - models are approximations, and discrepancies with reality are a source of uncertainty
- Experimental uncertainty – measurement error in live data
- Interpolation/extrapolation uncertainty – areas where data cannot be collected

Traditional statistical uncertainty quantification has focused on statistical methods based on data and statistical models (i.e., focuses on experimental, interpolation, and extrapolation uncertainty). New methods allow testers to combine those historical calculations with uncertainties in differences between models and live data. These methods will be critical for making the use of models and simulations in evaluations defensible in the future.

Sequential/Adaptive designs – Sequential and/or adaptive test designs are not a new concept. The Sequential Probability Ratio Test[11] is probably the most widely known in defense applications. However, in the armor testing community, sequential designs are common practice, and recent research has made testing more efficient (for example, see Three-Phase

---

[8] C. J. Wu, Post-Fisherian experimentation: from physical to virtual, Journal of the American Statistical Association 110 (2015) 612-620.

[9] Models and Simulations, 2002, http://www.dote.osd.mil/guidance.html.

[10] National Academy of Sciences Report (ISBN 0-309-06551-8), "Statistics, Testing, and Defense Acquisition, New Approaches and Methodological Improvements," 2012.

[11] Wald, A., & Wolfowitz, J. (1948). Optimum character of the sequential probability ratio test. The Annals of Mathematical Statistics, 326-339.

Optimal Design by Jeff Wu).[12] I expect many advances in this area, as the ability to update test designs based on previous data is both intuitively and mathematically appealing in both the Bayesian and frequentist paradigms.

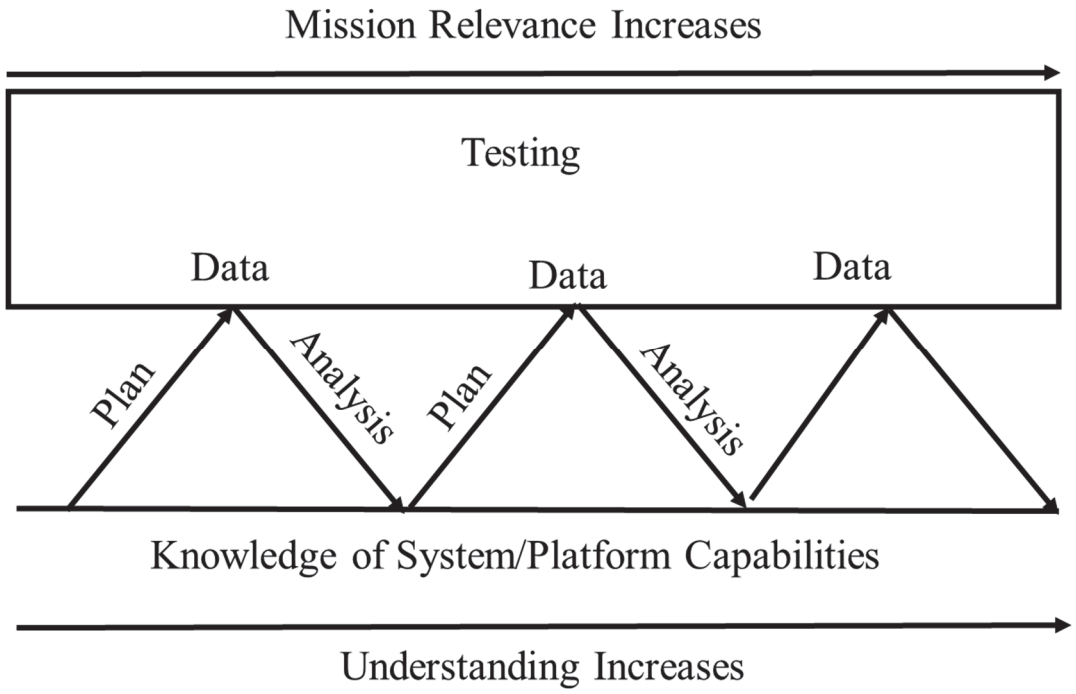

Figure 1. The iterative, sequential nature of test and evaluation, adopted from Box et al. (1978)

The statistical tools and methods I focused on are best used in an iterative, sequential process. Ideally, in T&E we should be "building" our case for evaluations using all available information. Figure 1 shows a progression of testing and analysis based over time. The idea is that when we start planning, we have limited information on the system. As a result, we potentially start by planning a space-filling design for a computer model or a sparse screening design of a sub-system to learn about important factors. As our knowledge increases and the system maturity increases, we can target our tests based on prior information and accumulate knowledge over multiple test phases. That building of knowledge is often reflected in Bayesian models, or models that capture growth (e.g., reliability growth tracking models).

Of course, we must always be mindful of what information we put into our models – as our analyses get more and more complex, we need to make sure that the T&E workforce includes individuals that understand the assumptions, risks, and limitations of these complex models. This will help ensure the right model is used for the right situation with the right knowledge. Because as the favorite statistical quote from Box goes, "All models are wrong, some are useful."

[12] Wu, C. J., & Tian, Y. (2014). Three-phase optimal design of sensitivity experiments. Journal of Statistical Planning and Inference, 149, 1-15.

However, we must also remember the corollary – models are very useful tools for conveying information!

Conclusion – A New Era for T&E

The future of T&E is promising. Advances in data science and statistics are revolutionizing our ability to incorporate subject matter expertise in planning tests and extract more useful information (and data) out of the tests we conduct. Moreover, it is easier than ever to publish and share tools. Gone (or nearly gone) are the days of passing around Excel files with simple calculations that can be corrupted. Code repositories and online applications make it possible for the T&E community to implement research from the academic community as soon as it is developed. At IDA, with the support of DOT&E, we have been working to capture these tools for use by the community (testscience.org). While still a work in progress, we hope to make it easy to share relevant tools broadly across the T&E community.

The next two issues of ITEA – Advanced Instrumentation and Information Systems Technology for T&E, and Statistical Methods in T&E – are perfectly suited to the future. I am always encouraged by the desire of the T&E community to keep learning about the systems we test and the test and analysis methods available. The way we collect, think about, and analyze data is rapidly changing, opening new doors for how we can improve our work as T&E practitioners.

Bio:

Dr. Laura Freeman is an Assistant Director of the Operational Evaluation Division at the Institute for Defense Analyses. In that position, she established and developed an interdisciplinary analytical team of statisticians, psychologists, and engineers to advance scientific approaches to DoD test and evaluation. Her focus areas include test design, statistical data analysis, modeling and simulation validation, human-system interactions, reliability analysis, software testing, and cybersecurity testing. Dr. Freeman currently leads a research task for the Chief Management Officer (CMO) aiming to reform DoD testing. She guides an interdisciplinary team in recommending changes and developing best practices. Reform initiatives include incorporating mission context early in the acquisition lifecycle, integrating all test activities, and improving data management processes.

During 2018, Dr. Freeman served as the acting Senior Technical Advisor for Director Operational Test and Evaluation (DOT&E). In that role, Dr. Freeman provided leadership, advice, and counsel to all personnel on technical aspects of testing military systems. She served as a liaison with Service technical advisors, General Officers, and members of the Senior Executive Service on key technical issues. She reviewed test strategies, plans, and reports from all systems on DOT&E oversight.

Dr. Freeman is the recipient of the 2017 IDA Goodpaster Award for Excellence in Research and the 2013 International Test and Evaluation Association (ITEA) Junior Achiever Award. She is a member of the American Statistical Association, the American Society for Quality, the International Statistical Engineering Association, and ITEA. She serves on the editorial boards for Quality Engineering, Quality Reliability Engineering International, and the ITEA Journal.

Dr. Freeman has a B.S. in Aerospace Engineering, a M.S. in Statistics and a Ph.D. in Statistics, all from Virginia Tech. Her Ph.D. research was on design and analysis of experiments for reliability data.