The Role of Test and Evaluation in Intelligence Community Sponsored Social and Behavioral Science Research

Jason Spitaletta, Ph.D.[1]; Ariel Greenberg, Nathan Bos, Ph.D.; Jonathon Kopecky, Ph.D.
The Johns Hopkins University Applied Physics Laboratory

Abstract

Ecologically valid social, behavioral, and neuroscience research is necessary to meet scientific and technological requirements of the intelligence community; particularly so when the requirement involved analytic tradecraft. Given the ongoing challenges of trusting the results of behavioral sciences given problems ranging from replication crises to publication bias to scientific fraud to over-reliance on models, (Earp & Trafimow, 2015; Maxwell, Lau, & Howard, 2015; Romer, 2016) this becomes an increasingly important component within applied research. The Intelligence Community (IC) invests in a wide variety of research ranging from short term solution-focused efforts support a particularly agency or subordinate to component to high-risk, high-payoff research programs to tackle some of the most difficult challenges of the agencies and disciplines across the IC. Many of those challenges entail social, behavioral, and neuroscience research, which present unique challenges when attempting to transition findings to the IC. IARPA employees a rigorous Test & Evaluation (T&E) process that seeks to ensure performer findings are both internally and externally as well as ecologically valid. This paper outlines some of the challenges and solutions developed by The Johns Hopkins University Applied Physics Laboratory (JHU-APL) in planning and executing T&E within the social, behavioral, and neurosciences.

Introduction/Statement of the problem

The Intelligence Community (IC) invests in a wide variety of research ranging from short term solution-focused efforts supporting a particularly agency to high-risk research programs to tackle some of the most difficult challenges amongst the various intelligence disciplines across the IC. The former tend to be funded by intelligence agencies themselves while the latter is the purview of the Intelligence Advanced Research Projects Activity (IARPA). Many intelligence research capability gaps entail challenges that might be addressed through social, behavioral, and neuroscience research, which present unique challenges when transitioning findings to the IC. Ecologically valid social, behavioral, and neuroscience research is necessary to meet scientific and technological requirements of the intelligence community; particularly so when the requirement involved analytic tradecraft. The social, behavioral, and neurosciences can provide insight into not only what needs to be analyzed but how it should be analyzed. In order for social, behavioral, and neuroscience to have the most operational impact, a robust Test & Evaluation (T&E) process should be integrated into US government sponsored research supporting the IC so that the products of said research and development can be ultimately assimilated in the most effective manner practicable.

Given the ongoing challenge of trusting the results of behavioral sciences given problems ranging from replication crises to publication bias to scientific fraud to over-reliance on models, (Earp & Trafimow, 2015; Maxwell, Lau, & Howard, 2015; Romer, 2016) T&E becomes an

[1] Corresponding author: Jason.Spitaletta@jhuapl.edu; Jason.A.Spitaletta@coe.ic.gov

increasingly critical component within applied research. This importance is relevant to not only how the US evaluates foreign scientific and technological developments but also how the IC incorporates those developments into its own policies and processes. Another, perhaps related, issue is the limited scientific advances made in the social sciences despite increased government funding (Watts, 2017). A suggestion is to focus interdisciplinary research on solving a particular social problem (Watts, 2017) and operational challenges faced by the intelligence community may meet some of those criteria. Solving, or even addressing, operational challenges is rarely simple or straightforward. There are often multiple scientifically valid suggestions about how to do so and thus it is incumbent on the funding organization to incorporate a means by which those suggestions can be independently validated and/or verified. IARPA, for example, employs a rigorous T&E process that seeks to ensure performer findings are both internally and externally as well as ecologically valid and this carries with it a number of scientific and operational challenges. The T&E process must be transparent to not only performers but also transition partners so all agree the process is rigorous, fair, and operationally relevant. This may entail selective replication, complex statistical methods, and/or sophisticated modeling and simulations approaches when the research cannot be scaled sufficiently such as when trying to understand or simulate societal-level social phenomena.

Challenges

T&E serves a number of purposes not the least of which is to ensure the IC that the US government funded research meets the requirements set forth by the government and does so in a manner that is rigorous, fair, and productive. This entails numerous challenges, key among them are transparency and validity.

Transparency

A sound T&E approach requires articulating evaluation criteria to performers (those funded directly by IC to perform a particular function). When and where possible the evaluation criteria should be included in the initial request for proposals so that the applicants can structure their responses to address the particular issue at hand. This also ensures a more focused peer-review process so those evaluating the submissions can compare the proposed methods with the established evaluation criteria.

The development of evaluation criteria must not only meet scientific standards but also those of the transition partner (end-user community). This can be challenging when the intent of a program is to extend scientific understanding of a problem relevant to the IC and, if a discovery is made then apply it. If the IC has not considered the application of an advanced concept or method, establishing a requirement can be somewhat conjectural. Nevertheless, the input of the end-user is vital in establishing how a particular approach should be evaluated.

There are additional, often legal, issues associated with evaluation criteria. Often it may require sharing intellectual property including raw data, code for a particular model, or proprietary algorithms. These types of disclosures are often necessary to ensure a tool, technique, or combination thereof works across the range of applications. This process is also necessary, particularly for software projects, to ensure the system is compliant and ultimately certified for use on the appropriate networks.

Validity

Effective T&E requires not only measures that are aligned with the objective, but also assurance that the measures are valid. Validity is an enduring concern when establishing evaluation criteria, particularly as intelligence constraints and lack of information can result in a tendency to assume a causal relationship due to sequential occurrence. Intelligence assessments

can be based on correlational data since an explicit causal relationship between friendly actions cannot be definitely linked to a target audience behavior due to the existence of innumerable extraneous and mediating variables. In T&E, particularly if using an experimental paradigm or modeling & simulation, the true state of the world is known and thus there is a "correct" or "incorrect" answer. While these luxuries don't exist in the real world, this artificiality is often necessary to evaluate analytical tradecraft or the tools used to conduct it.

Validity is arguably the most vital component of the T&E process, particularly when it entails human subjects research that seeks to understand intelligence personnel, their tradecraft, and/or their requirements. The study design must not only be scientifically sound but also operationally meaningful to the IC. There are numerous threats to internal validity including sampling bias (subject selection, statistical regression to the mean), attrition (subjects leaving a longitudinal study early, mortality in clinical trials), improper measurement (ineffective tools, poor technique), and/or artifact (response bias) (Kazdin, 2003). Threats to external validity can arise from the lack of experimental control of the process and include measurement unreliability sampling bias, and artifact (subjects being targeted through multiple means) (Kazdin, 2003). Perhaps the most important aspect of T&E experimental design is ecological validity or the degree to which the findings are applicable to a particular domain (Kazdin, 2003). Ecological validity often presents daunting challenges for T&E ranging from recruiting individuals who already work long, stressful hours, incorporating sufficient realism into an experiment yet doing so in an unclassified environment under Institutional Review Board (IRB) oversight, and/or verifying that the original requirements established in the evaluation criteria are 1) still applicable and 2) verifying they've been met.

The replication crisis in psychological science presents unique challenges for T&E; particularly when deciding whether to conduct a pure replication of performer research or a reproduction that varies experimental parameters, sample populations, and/or domains. Often the decisions entail not only scientific criteria but also operational and most assuredly financial. The concern about not only the repeatability but also the robustness of performer findings drives T&E experimental design. If the performer findings have small effect sizes or low statistical power, there is a desire to confirm those results. However, that desire is weighed against other performer findings, the generalizability of the findings, and the applicability of the performer experimental designs to the end-user requirements. While the decisions vary depending on the particular program, the process by which those decisions come about tend to focus on the aforementioned criteria.

Transition

Transition is the process of integrating performer findings, T&E validation and verification, and/or policy guidance to the operational components of the IC. Not every IC research effort need be transitioned in the near term to be successful. For example, IARPA's Knowledge Representation in Neural Systems (KRNS)[2] is more basic than applied science and, while it has resulted in interesting scientific discoveries and multiple scholarly articles in peer-reviewed journals it is not necessarily operationalized. In programs with more exploratory research objectives, T&E serves to validate and/or extend some of the initial performer research while also reporting on the contributions made to the state of the art. For other programs, for example IARPA's Sirius[3] and Crowdsourcing Evidence, Argumentation, Thinking and

[2] https://www.iarpa.gov/index.php/research-programs/krns
[3] https://www.iarpa.gov/index.php/research-programs/sirius?highlight=WyJzaXJpdXMiXQ==

Evaluation (CREATE)[4] transition takes on greater importance and thus should be a consideration from the creation of the Broad Area Announcement (BAA). Transition is not necessarily something that can be done well at later stages of a research program, particularly if the end-user community was not sufficiently involved in the process.

Recommendations for Future Research

Two areas in particular require ecologically valid T&E components that serve as a translational research bridge between the laboratory and the field; analytic tradecraft and credibility assessment. The National Institutes of Health (NIH) consider translational research the process of taking laboratory or preclinical research and developing appropriate clinical trials as well as research focused on improving practices derived from empirically-based approaches (Rubio et al., 2010). Translational research should also include nonclinical applications, such as the implementation of empirically-based approaches to intelligence tradecraft.

Example: Analytic Tradecraft

Intelligence analytic tradecraft has received increased attention in recent years and continues to be a research focus area for the IC. The Office of the Director of National Intelligence (ODNI)'s release of Intelligence Community Directive 203 *Analytic Standards* is a formal distillation and application of some of that research. Developing either tools or techniques for intelligence analysts and validating them under experimental controls entails not only the aforementioned validity concerns but also logistical considerations. The validity of performer experimental designs that do not use qualified intelligence analyst are often questions, but in some cases undergraduate cohorts (particularly those in majors at institutions from which the IC has historically recruited) are indeed suitable proxies for less available professionals.

It is difficult to replicate the Mental Effort Load, Time Load, and Psychological Stress Load (Reid & Nygren, 2001) that intelligence analysts experience while performing their tasks for any number of logistical and ethical reasons. Analysts experience a variety of stressors not the least of which is the knowledge that a mistake can cost American lives. It is challenging to replicate the reality of an intelligence analyst in a laboratory setting that seeks to test a tool or a technique and testing on analyst themselves requires the sample population to spend time not performing their primary function. Therefore, T&E must ensure analysts are indeed necessary as the sample population to answer specific research questions and that the tools, techniques, or combination thereof have been sufficiently tested in formative settings such that a summative test with professionals is justified.

While the ideal test of a new capability (tool, technique, or combination thereof) is to use actual intelligence analysts on actual intelligence data, this is often impractical and difficult to control. However, to accurately replicate not only the stressor of an intelligence analyst but also the complexity of the tasks required ranging from long-term quantitative estimates to short-term qualitative descriptions (Gerliczy, 2016) in an experimental context may require synthetic yet realistic data. Traditionally, the technique of agent-based modeling and simulation of social systems is used for hypothesis testing, forecasting, and sensitivity analysis. Another set of applications that is less explored involves gap identification – in data collection, in sample population, and in the theoretical underpinning of agent design. For example, (Nelson, Kennedy, & Greenberg, 2015) describe a method to unify microdata sources disparate in sample, time, location, fidelity, and topic to produce rich, synthetic, statistically-reasonable agents. This microdata synthesis approach reveals data collection gaps that if satisfied by focused interview

[4] https://www.iarpa.gov/index.php/research-programs/create?highlight=WyJjcmVhdGUiXQ==

or polling efforts would allow the incorporation of otherwise-orphaned microdata from one-off studies. Agents instantiated according to the architecture proposed by Epstein (2016) for neurocognitive realism and fed by the data synthesized in the fashion described above furnish an experimental environment unprecedented in verisimilitude. Such an environment also facilitates the identification of both data needs and theory weaknesses to guide further research endeavors.

Example: Credibility Assessment

Credibility assessment remains one of primary capability gaps in the IC and is thus an ongoing research topic of interest. Credibility assessment tools and techniques are required for initial employee screening as well as a variety of human intelligence (HUMINT) tasks including source operations, interrogation, and debriefing (Happel et al., 2015; Wolmetz, et al., 2015). Credibility assessment research is plagued by a lack of ecological validity, from overreliance on the mock crime paradigm to insufficiently trained participants, there are a number of areas that require improvement (Happel et al., 2015; Wolmetz, et al., 2015).

The effectiveness of credibility assessment, whether aided by technological means or not, is highly dependent on the interviewer as well as the interviewee; and too few studies treat this dyad as atomic, or inextricable without extensive parameterization. The High-Value Detainee Interrogation Group (HIG) (2016) has funded numerous social and behavioral science approaches to rapport-based interrogation, some of which have resulted in techniques that have been translated into tactics trained to US interrogators. Yet, the field requires evaluative frameworks directed to overcome the challenge of dyadic atomicity, and to apply the framework in a set of ecologically-valid experiments. In the wake of the controversial detention program, he US has invested in developing better dyadic approaches to credibility assessment yet many have not been subjected to T&E.

Framework-driven validation and verification experiments will enable assessment of effectiveness of the rapport-building approaches that aim to overcome resistance to and/or noncompliance with questioning approaches. Importantly, the experiments will be designed to also identify advantageous combinations of dispositional strengths and weakness of both interviewer and interviewee. While HUMINT related can be perceived as controversial (Borum,2006), the IC and Department of Defense (DoD) have become overly reliant on technological means of collecting intelligence at the expense of HUMINT (Kaminski, 2011) and thus more resources need to be dedicated to not only developing tactically sound techniques for gaining compliance and/or educing information but also developing evaluative frameworks to ensure those techniques are appropriately understood, trained, and ultimately applied.

Summary

The social, behavioral, and neurosciences have much to contribute to the IC and, as such, research funding for these sciences should continue. Those contributions may take the form of short-term, solution-focused applied research or longer-term higher-risk basic science. The IC must strike the appropriate balance between current and future requirements, often a daunting challenge for, to paraphrase Schopenhauer; "*talent hits a target no one else can hit; genius hits a target no one else can see.*" The IC must seek to understand and solve problems that require talent and to identify those that require genius. To ensure that funding maximizes benefits to the user community, a robust T&E component should be included along with that funding. That T&E component must remain transparent while seeking to verify end-user requirements using scientifically rigorous yet operationally relevant methods.

References

Earp, B. D., & Trafimow, D. (2015). Replication, falsification, and the crisis of confidence in social psychology. *Frontiers in psychology*, *6*, 621-632.

Epstein, J. (2016). Lecture on *Agent_Zero* and Generative Social Science. Summit on Social and Behavioral Sciences for National Security. National Academies of Science, Washington, DC.

Gerliczy, G. (2016). Lecture on the Role of Social and Behavioral Science in Intelligence Analysis. Summit on Social and Behavioral Sciences for National Security. National Academies of Science, Washington, DC.

Happel, M.D., Spitaletta, J.A., Hwang, G.W., & Wolmetz, M. (2015). *Detecting Deception and Concealed Information: Evaluation of the State of the Practice (REDD-2015-532).* Laurel, MD: The Johns Hopkins University-Applied Physics Laboratory.

High-Value Detainee Interrogation Group. (2016). *Interrogation Best Practices*. https://www.fbi.gov/file-repository/hig-report-august-2016.pdf/view. Accessed 2/14/16.

Howells, D.C. (2006). *Statistical Methods for Psychology, 6th Ed*. Belmont, CA: Wadsworth Publishing Company.

Kaminski, P. (2011). *Report of the Defense Science Board Task Force on Defense Intelligence Counterinsurgency (COIN) Intelligence, Surveillance, and Reconnaissance (ISR) Operations*. Washington, DC: Office of the Under Secretary of Defense for Acquisition, Technology and Logistics.

Kazdin, A.E. (2003). *Research Design in Clinical Psychology, 4th Ed*. Boston: Allyn and Bacon.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? What does "failure to replicate" really mean?. *American Psychologist*, *70*(6), 487-498.

Nelson, J. B., Kennedy, W. G., & Greenberg, A. M. (2015). Agents and Decision Trees from Microdata. In *Proceedings of the 24th Conference on Behavior Representation in Modeling and Simulation (BRIMS)*.

Open Science Collaboration. (2012). An open, large-scale, collaborative effort to estimate the reproducibility of psychological science. *Perspectives on Psychological Science*, *7*(6), 657-660.

Reid, G. B., & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in psychology*, *52*, 185-218.

Romer, P. (2016). The trouble with macroeconomics. *The American Economist*. http://ccl.yale.edu/sites/default/files/files/The%20Trouble%20with%20Macroeconomics.pdf. Accessed 2/14/16

Rubio, D. M., Schoenbaum, E. E., Lee, L. S., Schteingart, D. E., Marantz, P. R., Anderson, K. E., ..& Esposito, K. (2010). Defining translational research: implications for training. *Academic medicine: journal of the Association of American Medical Colleges*, *85*(3), 470-475.

Watts, D. J. (2017). Should social science be more solution-oriented?. *Nature Human Behaviour*, *1*, 0015.

Wolmetz, M., Pohlmeyer, E.A., Spitaletta, J.A., Scholl, C.A., Greenberg, A.M., Hwang, G.W., & Happel, M.D (2015). *Detecting Deception and Concealed Information: Evaluation of the State of the Art (REDD-2015-533).* Laurel, MD: The Johns Hopkins University-Applied Physics Laboratory.