ARL-MR-1144 ● JAN 2026

Data Quality Assurance for Test and Evaluation of Complex Soldier–Machine Formations

by Jason S. Metcalfe

NOTICES

Disclaimers

ARL-MR-1144 ● JAN 2026

Data Quality Assurance for Test and Evaluation of Complex Soldier–Machine Formations

Jason S. Metcalfe
DEVCOM Army Research Laboratory

REPORT DOCUMENTATION PAGE

<table>
<tr><td rowspan="2">1. REPORT DATE
January 2026</td><td rowspan="2">2. REPORT TYPE
Memorandum Report</td><td colspan="2">3. DATES COVERED</td></tr>
<tr><td>START DATE
1/12/2023</td><td>END DATE
8/22/2025</td></tr>
<tr><td colspan="4">4. TITLE AND SUBTITLE
Data Quality Assurance for Test and Evaluation of Complex Soldier–Machine Formations</td></tr>
<tr><td>5a. CONTRACT NUMBER</td><td>5b. GRANT NUMBER</td><td colspan="2">5c. PROGRAM ELEMENT NUMBER</td></tr>
<tr><td>5d. PROJECT NUMBER</td><td>5e. TASK NUMBER</td><td colspan="2">5f. WORK UNIT NUMBER</td></tr>
<tr><td colspan="4">6. AUTHOR(S)
Jason S. Metcalfe</td></tr>
<tr><td colspan="3">7. PERFORMING ORGANIZATION NAME(S) AND ADDRESS(ES)
DEVCOM Army Research Laboratory
ATTN: FCDD-RLA-FB
Aberdeen Proving Ground, MD 21005</td><td>8. PERFORMING ORGANIZATION REPORT NUMBER
ARL-MR-1144</td></tr>
<tr><td colspan="2">9. SPONSORING/MONITORING AGENCY NAME(S) AND ADDRESS(ES)</td><td>10. SPONSOR/MONITOR'S ACRONYM(S)</td><td>11. SPONSOR/MONITOR'S REPORT NUMBER(S)</td></tr>
<tr><td colspan="4">12. DISTRIBUTION/AVAILABILITY STATEMENT
DISTRIBUTION STATEMENT A. Approved for public release: distribution unlimited</td></tr>
<tr><td colspan="4">13. SUPPLEMENTARY NOTES
ORCID: Jason S. Metcalfe, 0000-0001-9086-9962</td></tr>
<tr><td colspan="4">14. ABSTRACT
This report details the Data Quality Assurance (DQA) processes developed and implemented within ARL’s Human–Autonomy Teaming program during FY23 and evolving through the end of FY25. Recognizing the critical importance of data integrity in Soldier–Machine Integration research, this work outlines a framework for continuous DQA interwoven throughout the entire experimental lifecycle. The report emphasizes a Directed Action Team structure, aligning specialized expertise with specific phases of experimentation from protocol development and measurement definition to data collection and analysis. Key best practices are identified, including proactive protocol management, robust communication strategies, and iterative tool improvement. This report provides practical guidance for Science and Technology Reinvention Laboratories seeking to establish and sustain high-quality data pipelines, ultimately enhancing the reliability and validity of research findings and accelerating the transition of innovative technologies to the Warfighter. The lessons learned underscore the necessity of balancing rigorous planning with adaptability in dynamic research environments.</td></tr>
<tr><td colspan="4">15. SUBJECT TERMS
Data Quality Assurance, Science and Technology Reinvention Laboratory, STRL, Army Research Laboratory, ARL, Human–Autonomy Teaming, HAT, Data Management, Document Control, Data Validation, Technology Demonstration, Demonstration Experiment, Humans in Complex Systems</td></tr>
<tr><td colspan="3">16. SECURITY CLASSIFICATION OF:</td><td rowspan="2">17. LIMITATION OF ABSTRACT
UU</td><td rowspan="2">18. NUMBER OF PAGES
53</td></tr>
<tr><td>a. REPORT
UNCLASSIFIED</td><td>b. ABSTRACT
UNCLASSIFIED</td><td>C. THIS PAGE
UNCLASSIFIED</td></tr>
<tr><td colspan="3">19a. NAME OF RESPONSIBLE PERSON
Jason S. Metcalfe</td><td colspan="2">19b. PHONE NUMBER (Include area code)
(520) 691-6430</td></tr>
</table>

STANDARD FORM 298 (REV. 5/2020)
Prescribed by ANSI Std. Z39.18

Contents

List of Figures

Acknowledgments

I gratefully acknowledge, first and foremost, the 112+ active-duty Soldier participants for offering their full attention and complete engagement throughout our joint technology excursions; their units are listed below. Further, we share deep appreciation for the 61+ Army civilian leaders, scientists, and engineers who worked alongside 44+ skilled experts from contractor organizations, all of whom were instrumental in the development, evolution, and rigorous testing of the ARL Information for Mixed Squads (INFORMS) laboratory at Aberdeen Proving Ground, Maryland.

We acknowledge and credit the use of multiple (commercial and Army-owned) technology-based "intelligent assistants" for conducting and consolidating the work of the Data Quality Assurance (DQA) team, as well as for abstracting over a thousand pages of appropriately redacted meeting transcripts, notes, and incremental technical products into this generalized report.

With gratitude I recognize the following Warfighter units (in sequence of demonstration event year):

Soldier, Operator, Maintainer, and Evaluation Program, Aberdeen Test Center

34th Infantry Division, Minnesota Army National Guard

Army Evaluation Center under the Army Test and Evaluation Command

3rd Armored Brigade Combat Team, 1st Armored Division

1st Armored Brigade Combat Team, 1st Cavalry Division

1st Armored Brigade Combat Team, 1st Infantry Division

3rd Armored Brigade Combat Team, 1st Cavalry Division

1st Battalion, 7th Field Army, 1st Infantry Division

1st Squadron, 1st Cavalry Regiment and 2nd Brigade, 1st Armored Division

Contributors to ARL INFORMS between 2018 and 2025 include:

DEVCOM ARL

Baker, Anthony
Brewer, Ralph W. II
Canady, Jonroy
Carter, Evan
Cesar-Tondreau, Brian
Chhan, David
Cohen-Hoffing, Russell
Cooper, Laura
Cox, Katherine
Crowell, Philip (Harrison III)
DeCostanza, Arwen
Ericson, Mark
Evans, Bill (Arthur William III)
Files, Benjamin
Fink, Isadora
Fitzhugh, Sean
Fluitt, Kim
Forster, Daniel
Gaston, Jeremy
Giammanco, Cheryl
Gremillion, Gregory
Gutstein, Steve
Hansberger, Jeffrey
Hairston, W. David
Hill, Susan
Holder, Eric
Johnson, Craig
Kerick, Scott
Krausman, Andrea
Lakhmani, Shan
Larkin, Gabriella
Loukota, Patrick
Lieberman, Gregory
Marathe, Amar
McDowell, Kaleb
Mermagen, Tim
Metcalfe, Jason
Napier, Samantha
Neubauer, Catherine
Perelman, Brandon
Pollard, Kimberly
Rexwinkle, Joe
Riggs, Paul
Roy, Heather
Schaefer-Lay, Kristin
Scharine, Angelique
Schizas, Ioannis
Shorter, Paul
Sinks, Chris
Smith, Thomas
Story, Brittany
Thurman, Steven
Wright, Julia
Yu, Alfred

DEVCOM Ground Vehicle Systems Center

Berati, Ben
Haynes, Benjamin
Kowachek, Dave
Mikulski, Christopher
Mikulski, Dariusz
Paul, Victor
Tierney, Terrance

DCS Corp

Atwater, Corey
Bialek, Peter
Boutros, Yanall
Bowers, Travis
Byrnes, Johnny
Campanelli, Joe
Cannon, Marcel
Choi, John
Chou, Patrick
Crites, Courtney
DeAngelis, Evan
Diego, Richard
Dinare, Mesgana
Egland, Ian
Fernandez, Robert
Fettig, Claire
Gordon, Stephen
Grant, Lauren
Groves, Paul
Jagielski, Justin
Jaswa, Matthew
Johnson, Tony
Kellihan, Bret
Lee, Timothy
Majeed, Aamir
Majeed, Aarif
Maslin, William
Meli, Ben
McBrearty, Mike
McGhee, Sean
Montana, Poole
Nham, Xung
Nonte, Mike
Owen, James
Protzman, William
Rohaly, Thomas
Scheiner, DCS
Siero, Matt
Song, Jae
Sparr, Henry
Tisdale, Madeline
Walker, Anthony

D-Prime, LLC

Brooks, Justin
Kucukosmanoglu, Murat

Executive Summary

Established in 2018, the U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) Information for Mixed Squads (INFORMS) laboratory at Aberdeen Proving Ground, Maryland, expanded and then reestablished in 2021 following the global pandemic. This facility (Figure ES-1) is world class in its unique capabilities and, more importantly, is a remarkable achievement in consistent delivery of complex, realistic, immersive operational simulations to active U.S. Soldiers who have deployed similar systems in theater. Complete with a well-practiced and organized Red Team (the "OPFOR"), the INFORMS is an exceptional environment, operated by a specialized community, that elicits authentic warfighting behaviors and Soldier team dynamics under emulated high-intensity operations. INFORMS is a testament to the dedication and expertise of all involved, as well as to the vision of the ARL leaders that championed the creation of this important Army resource.

Figure ES-1. One configuration of ARL's INFORMS Laboratory. The seven Soldier participants sitting in the wireframe simulator (left) are occupying an emulated future manned control vehicle (MCV) shown to the right (top). From front to back, a single vehicle commander sits in the lead position and gunner-driver pairs sit in the successive rearward rows to control the MCV, as well as two uncrewed robotic combat vehicles (right-bottom). Combined with a second wireframe simulator (not shown), doubling the simulated asset count to a section of six vehicles, this uniquely immersive mixed-reality simulator provided a high-fidelity operational space to safely explore advanced, yet experimental, next-generation systems in high-risk direct-combat scenarios, with appropriately trained and experienced operators.

The processes documented in this report are directly attributable to the challenges encountered and lessons learned by these teams and their forebears on whose work INFORMS was based: Cosenzo et al.[1]; Lance et al.[2]; Manteuffel et al.[3]; Metcalfe et al.[4]; Schaefer et al.[5]; and Vettel et al.[6]. Their collective commitment to delivering accurate, insightful analyses was foundational to the success of this experimental ecosystem and especially the processes for technology transition that emerged as both inspiration and future guides.

Finally, we acknowledge and credit the use of multiple (commercial and Army-owned) technology-based "intelligent assistants" for conducting and consolidating the work of the Data Quality Assurance (DQA) team, as well as for abstracting over a thousand pages of appropriately redacted meeting transcripts, notes, and incremental technical products into this generalized report. This report, therefore, incorporates output generated with the support of multiple intelligent text synthesis tools, commonly known today as artificial intelligence (AI) or large language models (LLMs).

- All AI-assisted content was guided, structured, and validated by the lead author.
- LLMs were used as tools for summarization, organization, language refinement, and image production.
- No content was auto-generated in full, or incorporated without human edits, revisions, as well as critical oversight for accuracy.

---

[1] Cosenzo K et al. Impact of 360° sensor information on vehicle commander performance. Army Research Laboratory (US); 2011 Mar. Report No.: ARL-TR-5471.

[2] Lance B, Vettel J, Paul V, Oie KS. The mission-based scenario: rationale and concepts to enhance S&T research design. Presented at National Defense Industrial Association Ground Vehicle Systems Engineering and Technology Symposium (NDIA-GVSETS); 2011 Aug 9–11; Dearborn, Michigan.

[3] Manteuffel C et al. A methodology for the assessment of 360° local area awareness displays. In: Guiell JJ, Bernier KL, Thomas JT, Desjardins DD, editors. Proceedings of the Display Technologies and Applications for Defense, Security, and Avionics V; and Enhanced and Synthetic Vision 2011. SPIE Defense, Security, and Sensing; 2011 Apr 25–29; Orlando, FL. Vol. 8042. SPIE; c2011. p. 126–139.

[4] Metcalfe JS et al. Experimentation and evaluation of threat detection and local area awareness using advanced computational technologies in a simulated military environment. In: Gerhart GR, Gage DW, Shoemaker CM, editors. Unmanned Systems Technology XII. SPIE Defense, Security, and Sensing; 2010 Apr 5–9; Orlando, FL. Vol. 7692. SPIE; c2010. p. 72–83.

[5] Schaefer KE et al. Assessing multi-agent human–autonomy teams: U.S. Army robotic wingman gunnery operations. Micro- and Nanotechnology Sensors, Systems, and Applications XI. SPIE Defense + Commercial Sensing; 2019 Apr 14–18; Baltimore, MD. Vol. 10982. SPIE; c2019. p. 353–372.

[6] Vettel JM et al. Mission-based scenario research: experimental design and analysis. Presented at the National Defense Industrial Association Ground Vehicle Systems Engineering Technology Symposium (NDIA-GVSETS); 2012 Aug 14–16; Warren, Michigan.

- Inputs provided to LLMs included previously authored materials (e.g., laboratory notes, and transcripts), closely reviewed, and redacted as needed for sensitivity.

All final outputs were subjected to rigorous human review through multiple expert volunteers independent from the DQA team.

1. Introduction

Cross-national competition has long fueled innovation across major social sectors, including commerce, sports, entertainment, politics, and especially defense. Today, the global environment exerts enormous pressure on the defense innovation ecosystem, which has historically functioned as a steadfast institution, and yet often lags in pacing both technical and social disruptions. Innovation now unfolds at such breathtaking speed that even the global 24-h news cycle struggles to keep pace with factual developments. Especially in this post–large language model (LLM), mass data-driven era, convergent breakthroughs in artificial intelligence (AI), cheap, deployable drones, personal devices and wearables, Internet of Things (IoT), context-aware networks, quantum-enhanced computation, and colossal data centers are reshaping the threats and opportunities that drive military innovation in lethality, protection, support, and sustainment. With unprecedented operational complexity, however, rapid prototyping and innovation under high risk is a constant expectation; legislative and social uncertainties only compound the complexities. The DoD must therefore ensure that both the informational integrity and the broad value contained within its vast data stores is maintained in accord with its import across the acquisition lifecycle. Such assurances often reside where the dataset is generated but may also be coordinated by stakeholder values and constraints.

Then enters the DoD Science and Technology Reinvention Laboratories (STRLs), which are uniquely chartered to bridge the gaps between early-stage research, practical applications, and operationally deployable products that address the most pressing current and future Warfighter needs. Unlike conventional laboratories that tend to focus on narrowly defined projects within tightly constrained silos, STRLs can address broad innovation goals through fundamental, integrative research with experiments that are selected and designed for rigorous, military-focused transition into capability offsetting, lifesaving, and conflict-ending technical advantages. Within STRL environments, experts from diverse fields like computer science, engineering, behavioral science, and beyond coalesce and translate foundational discoveries and principles into real-world capabilities. Whether refining an autonomous ground vehicle or analyzing next-generation intelligence flows, STRL experts must tackle messy realities of operational experimentation while manifesting novel scientific and technical value through all aspects of the Soldier selection, development, and readiness lifecycle.

Preserving the mission-critical value contained within the research, development, test, and evaluation (RDTE) data of focus at STRLs is therefore an essential strategic imperative. High-quality data not only underpin accurate performance metrics for advanced Soldier–AI formations but also can shape decisions that ripple through the entire capability lifecycle. The next section introduces STRLs in greater depth, explaining how their unique authorities, operational posture, and core philosophies make them vital conduits for Warfighter-focused innovation. The remaining portion of the report outlines the critical challenge of Data Quality Assurance (DQA) in these contexts while emphasizing the imperative to establish and secure the core stakeholder value through its best practices.

1.1 About STRLs

STRLs are specialized research entities within the DoD, which are granted unique management and hiring authorities under the Floyd D. Spence National Defense Authorization Act for Fiscal Year 2001 (Floyd D. Spence . . . 2000). Designed to accelerate transitions of emerging technologies from academic or basic research outcomes into fully operational capabilities, and to do so at a pace that out-competes peers globally, STRLs occupy a pivotal niche in the innovation pipeline of the DoD and the United States. In contrast to strict academic laboratories, STRLs pursue disruptive foundational science alongside and integrated with applied engineering, prototyping, and iterative experimentation that speak directly to Warfighter needs. Therefore, STRL scientists and engineers (S&Es) are de facto subject matter experts (SMEs) and advisors for national and military leadership to support complex decisions and facilitate the long-term strategic planning necessary to achieve continual improvement, broad resilience, and sustained advantage in our increasingly technical society.

Each STRL maintains a highly interdisciplinary and competitive workforce, relying on foundational research scientists working alongside talented specialists from most domains of modern engineering and technology integration. These professional S&Es operate in world-class, Defense-specialized facilities, co-located within force-based Command elements and other major partner DoD installation locations throughout the United States and the world. This tight integration enables bi-directional interfacing of STRL S&Es with strategists and end-user Soldier communities, and this provides essential feedback loops that enable the direct dialogue needed for facilitating the translation of foundational science theory into effective, asymmetric battlefield advantages.

In many cases, STRLs leverage their SMEs to forge strong, productive partnerships across the broader science and technology (S&T) ecosystem, engaging colleges and universities, public and private research institutions, and a vast network of industry and commercial entities capable of scaling scientific breakthroughs to real-world applications. Many STRLs exercise special individual flexibilities because of their ability to connect (and/or compete) with and integrate throughout the vast world of scholarship, innovation, and business communities. These flexibilities enable the recruitment of the very best talent to build and maintain a robust capability base inside the DoD, which is essential for innovation capabilities that cross into protected informational spaces. This agility is especially critical in the context of modern development and product lifecycles, complete with increasingly frequent market disruptions and unexpected events that demand swift, adept RDTE pivots in cooperation with intellectual and industry elites.

Beyond their advantages of structure and access, STRLs are demonstration-focused, end-user evaluated ecosystems that commit to the highest possible levels of operationally relevant realism, intentionally cultivating and managing the kinds of complexities that other laboratories spend millions to simply avoid. Rather than confining new concepts to small-scale, controlled studies, STRLs regularly expose prototypes and methodologies to austere environments and multifaceted missions that mimic real-world complexity. These regular technology exercises, Soldier touchpoints, and advanced demonstrations often involve live exercises with authentic, ruggedized systems, deployed in Soldier–machine integrated and localized scenarios or distributed national and multinational collaboration events.

STRLs must directly accept the need to establish robust DQA processes and methods as central to their respective visions of success. Indeed, through their novel accesses and capacities, STRLs are poised to turn every sensor readout, human-factors evaluation, or system certification event into evidence that shapes strategic acquisition decisions and global outcomes. Consequently, STRLs are uniquely poised to conduct and verify cutting-edge research that can evolve into reliable, mission-ready capability, reinforcing the essential aim of Army innovation: to deliver decisive overmatch through operationalized Warfighter know-how.

1.2 STRLs Are Engines Designed to Transform Data to Value

Building on the capacity for advanced prototyping, immersive experimentation, and live, human-in-the-loop demonstrations, STRLs stand out for their ability to transform raw observations into valuable, actionable, Warfighter-focused knowledge and strategic advantage. Each trial or experiment generates data that must be evaluated through a series of metrics at various levels of maturity, and ultimately demonstrate technical feasibility, operational relevance, and cost-efficiency. By integrating engineering metrics like reliability and throughput with contextual factors like human–machine team effectiveness or environmental constraints, these STRL environments can ensure datasets are scrutinized through a lens of real-world effectiveness and impact, rather than focal, commercial wins.

This process of synthesis hinges on cross-functional coordination among specialized experts who have cultivated and refined a unique knowledge of how their insights, skills, and experiences apply to the defense and security of the nation. In many instances, operational stakeholders join S&Es in real-time technology excursions to refine objectives, pivot test parameters, or explore emergent behaviors. The net effect is an iterative feedback loop, where experiments are not static, one-off evaluations but dynamic events continually shaped by on-the-spot discoveries and annual strategic shifts in research needs. Through structured data governance and checkpoints as needed, STRLs must fuse diverse inputs into unique value propositions that directly inform acquisition milestones, doctrinal adjustments, and larger strategic considerations.

At the heart of this conversion from data to value lies an implicit and necessary commitment to DQA, which bears an essential onus for vigilant attention to evolving knowledge management throughout a complex acquisition ecosystem. These laboratories benefit most when they standardize collection protocols, conduct assessments in accord with rigorous plans, and apply transparent methods toward capturing rich context under clear test and evaluation objectives. To contribute value over multiple cycles of testing, such measures must accumulate toward a robust digital corpus, enabling decision-makers to trust the insights gleaned, even when mission parameters must shift or unexpected edge cases emerge. By viewing every test event as an opportunity for holistic learning, STRLs have a clear avenue to maintain the momentum and clarity to drive groundbreaking discovery into the systematic production of decisive, winning solutions for the DoD.

1.3 Report Overview

This report builds on the above context to explore how S&Es at the U.S. Army Combat Capabilities Development Command (DEVCOM) Army Research Laboratory (ARL) (Figure 1), an STRL that specializes in foundational, early applied, and science-to-prototype RDTE programs, have harnessed a recurring demonstration experiment under their Human–Autonomy Teaming (HAT) Essential Research Program to create a robust, high-quality DQA process that supports Soldier–Technology Team Assessment for impactful innovation.

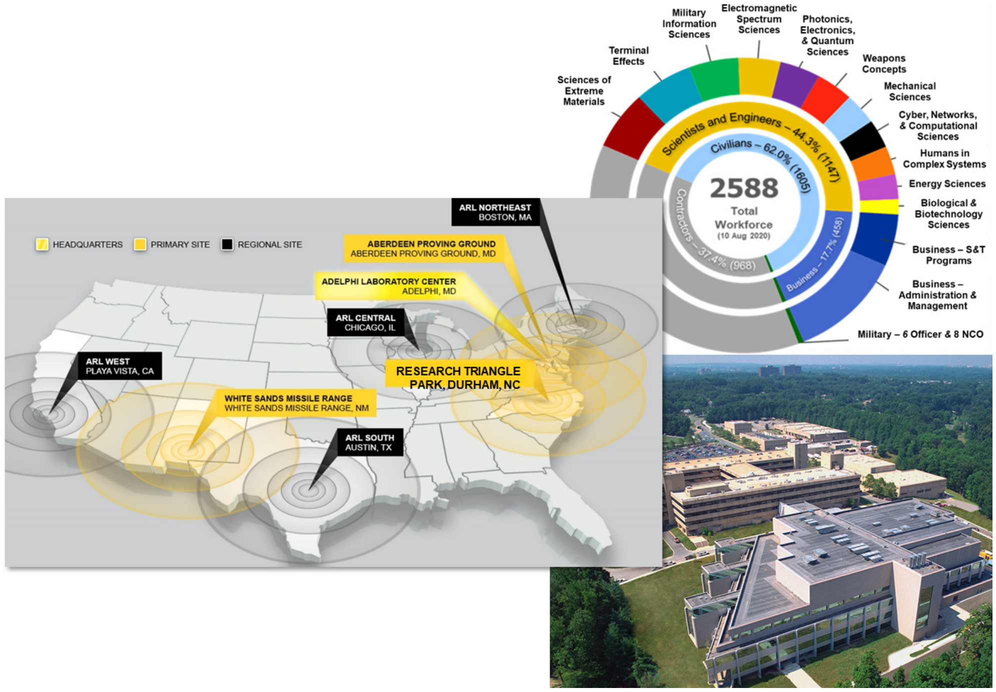

Figure 1. A visual summary of ARL's footprint and composition, including an exhibit of the Harold A. Zahl Laboratory at ARL Headquarters (Adelphi, Maryland). Strategically poised to develop national and regional partnerships throughout the continental United States and abroad, ARL collaborates across a vast range of foundational and applied scientific, engineering, and technological challenge areas of anticipated importance for near and far future Defense and Security operations.

Section 2 begins by orienting discussion onto the concept of data quality as value preservation, which is a focal construct around which the ARL HAT DQA process was built, illustrating how data and institutional objectives are expertly interwoven to create outcomes beneficial to the operational Army, the broader DoD, and other stakeholders. Section 3 describes and evaluates the strategies and tactics by which DQA may be aligned and coordinated in dynamic ecosystems, highlighting execution models and best practices that maintain data integrity under evolving mission requirements. Section 4 then examines how these strategies were translated into practice across a full demonstration cycle, detailing how the ARL team successfully aligned critical DQA tasks with shifting experimental priorities, personnel constraints, and stakeholder timelines. At the end of this report, readers should have a clear picture of how complex, operational RDTE data, laboratory protocols, and stakeholder collaboration can coalesce to forcefully drive forward-leaning, Warfighter-centric outcomes toward integration speeds that evolved to support modern innovation cycles in information-controlled environments.

2. Value, Data, and Their Relationship

Value can sometimes appear obvious, such as a single innovation that directly improves Soldier effectiveness and/or safety; as happened, for example, with the number of Soldier lives spared with the emergency introduction of the double v-hull in response to mass use of improvised explosive devices during operations in Iraq and Afghanistan (Good 2012). However, in many cases, value hides within massive datasets that encode myriad cross-domain interactions, or inside long-term trends, emerging only when data are thoughtfully curated, aggregated, and interpreted. Whether the aim is to enhance lethality, reduce costs, or accelerate readiness, it is the quality and context of data that ultimately shapes which valued outcomes may be achieved.

A helpful tool for visualizing the many forms and manifestations of value within a dataset is the **Value Compass**, introduced by Douglass Laney in 2022 (Laney 2022; Figure 2). The compass maps across two general axes, such as those polarized as *revenue versus costs* and *risks versus quality*, to create a conceptual reference for identifying the potential value in an available corpus of data or an impending technology exercise. In defense contexts, the specific individual types of value may correspond with critical factors like force readiness (experience, retention, and completeness), logistical efficiency (compliance and safety, efficiency, and stability), or mission adaptability (agility and differentiation). Here, Laney represents the general construct of **value** as multifaceted and covers a broad scope of potentially attractive outcomes, depending on the goal orientation of the stakeholder. For example, a program manager will see added value from maximizing quality, completeness, efficiency, and simplicity while minimizing risks and assuring compliance and safety, whereas a strategic planner and organizational executive will prioritize margin, revenue, growth, and reputation, among others, depending on their sector. In general, the value generated by *any* given research and innovation effort *will be judged by the ends that effort was to serve*; that is, if a new capability reduces inefficiencies and costs, which would be a clear benefit to the organizational bottom line, but it reduces safety, operational simplicity, and increases overall risk—opposite to the initial objective of the effort—then it would be judged as needing further development or design revision to restore and preserve the intended value of risk reduction for Soldiers.

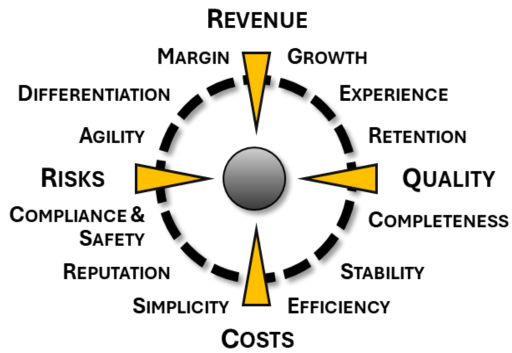

Figure 2. The Value Compass, as proposed in Laney (2022). Colors added by author.

It is incumbent upon all involved in the RDTE enterprise, from developers to decision-makers, to maintain this balanced and comprehensive view of *value* when developing processes and evaluations involving the outcomes and progress of our research and development investments. Here, we argue, for any given laboratory, objective, and technical exercise, data quality is defined by whether the intended value within the data may be realizable according to stakeholder preferred metrics. For us, DQA is synonymous with value curation, preservation, and protection.

2.1 Three Fundamental Premises

1) **Value drives decisions**. At its core, quality management enables the derivation of value from data (DAMA International 2024). Efforts in STRLs are likewise motivated by the pursuit of *defense-focused* value, whether expressed as enhanced operational capability, improved efficiency, reduced budget, or minimized risk. Such value is not an abstract ideal, but it is instead the practical outcome that matters most to our stakeholders, from program managers to senior leadership to Soldiers in the field. When identifying whether a new technology or protocol meets its objective, the ultimate question typically relates back to *What value is being advanced?* If value remains unclear or unarticulated, and if that value is not well-mapped to the data outcomes, then even the most sophisticated technical efforts will fail to inform meaningful decisions.
2) **Data provides the vehicle for realizing value in RDTE activities**. Although value provides the aim, it is **datasets** that are collected, shared, and interpreted across diverse contexts that ensure each initiative can move from concept to actionable result. In the Army, the acronym of DOTMLPF-P* suggests data may range broadly; from fuel, temperature and pressure sensor readings on equipment health, to Soldier communication

* Doctrine, Organization, Training, Materiel, Leadership, Education, Personnel, Facilities, and Policy

and system usage recordings, to individual behavioral or company-level operational analytics that point to readiness levels. Each of these data sources holds the potential to yield insights regarding specific improvements, but only if the right tools and frameworks are in place to orchestrate, verify, and meaningfully fuse them. Therefore, data is far more than raw input; quality data is a conduit through which latent value becomes visible and usable.

3) **Data quality preserves and protects value**. Central to all is the recognition that quality is essential for data to retain or amplify its inherent value. In an operationally complex environment, where mistakes in calibration, metadata capture, or timing can have disproportionate impacts, DQA becomes a critical safeguard. Well-defined protocols, systematic checks, and robust validation standards help ensure that data remains traceable, accurate, and complete throughout its lifecycle. When properly executed, DQA goes beyond error reduction toward *active curation of value*—allowing decision-makers to hold trust in its derived insights, make decisive choices, confidently take considered actions, and to modernize tools and tactics with adaptability and foresight.

2.2 Data Quality and Distributed Responsibility

Even the most elegant framework for managing data cannot succeed if data quality is treated as a specialized concern relegated to a single unit or a narrow subset of individuals. Data quality is neither a single checkpoint in the research pipeline nor is it an individual or small team of analysts, rather, data quality is the responsibility of all who claim a stake in the value it generates. *This value is socially constructed*, shaped by countless inputs and interactions among individuals, teams, and broader organizational systems. By adopting an ecosystem perspective, we recognize that every stakeholder—from the technical team to the Soldier gathering sensor data in the field, and through to the intelligence team analyzing results in the laboratory to the Commanding General making strategic decisions—will influence the future reliability and interpretability of their data.

Figure 3 underscores how multiple circles of context, each being larger and more inclusive than the previous, intersect and influence one another, reminding us that data's usefulness depends on cohesive, collaborative practices rather than isolated, one-off initiatives. Here, we borrow an explanatory *psychological construct* from human developmental science as an aid for conceptualizing the layered nature of value growth and manifestation through a research and development activity. We refer to Bronfenbrenner's (1979, 1989) *ecological theory of human development*, originally used to illustrate how a human (innermost **microsystem**) developing from child to adulthood is shaped by a complex *milieu* within which their lives are immersed, composed of many nested and interacting environmental subsystems from parents and siblings (**mesosystem**) to churches and schools (**exosystem**), and to the broadest layer of governance and international relations (**macrosystem**).

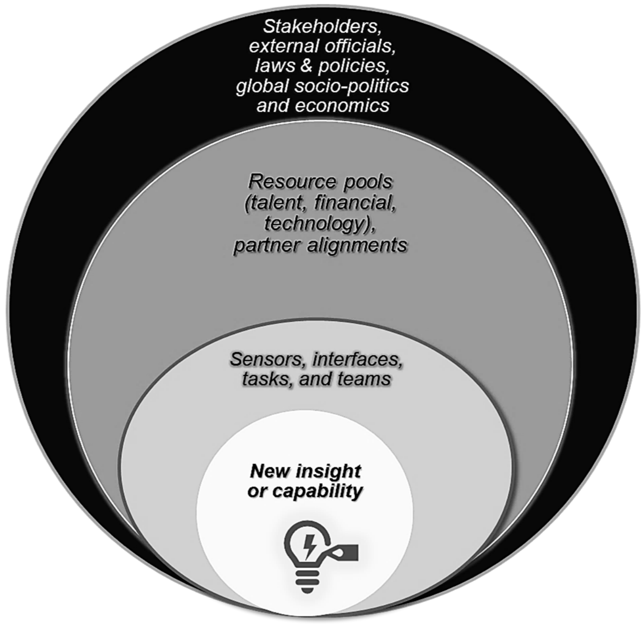

Figure 3. Layers of DQA as an ecosystem of nested goal-based subsystems that mutually influence and shape data quality through their complex interactions.

Similarly, we may conceive each data-driven endeavor in our projects as focused on fleshing out a new insight or capability to be cultivated and developed in a test or demonstration exercise. Like the young human who requires guidance and support, nurturance and social modeling from the surrounding ecosystem, a developing concept or capability (*micro*-) must be mindfully influenced under a simultaneous and complex mixture of direct (*meso*-) interactions among sensors, interfaces, tasks, and teams, intermediate (*exo*-) organizational dynamics spawned by resource management and directives from project and program fiduciaries, and, finally, larger (*macro*-) strategic or policy guidance that encodes stakeholder mandates and technology transition goals.

With Bronfenbrenner's analogy in mind, you may appreciate why no individual working from within an isolated role or office can be expected to maintain complete data integrity from start to end, especially if others neglect their responsibility and there is no systemically traceable accountability structure. As with the human's richness of capability, adaptation, and understanding, literally cultivated through a complex socio-biological ecosystem, so is the value contained within the flow of bits and bytes of the data that are captured during an experiment. Conversely, when each person acknowledges the communal onus for capture, curation, and safeguarding data and for collectively preserving value, DQA evolves from a standalone policy into an *integral cultural norm* that enables advanced science and technology concepts to rest on a foundation of trustworthy data.

2.3 Layered Ecosystem: Insight and Capability at the Center

If we regard each insight sought and capability objective as we might consider a person whose opportunity to thrive depends on us, then we may also envision the importance of our conscious management of the interactions and events that directly and indirectly affect overall ecosystem robustness. To yield significant return on investment and produce high-value outcomes, researchers at STRLs must transform desired capabilities and insights into analytically tractable questions. Doing this requires well-crafted internal logic, clarity of definition, and planful alignment with operational constraints. Much like how we nurture the young, our work recognizes the very way that we cautiously define and mindfully steward each capability sought through a process of:

- **Goal orientation** – articulating clear questions and objectives,
- **Metric derivation** – building testable hypotheses and objective measures,
- **Procedure cultivation** – vetting qualitative protocols,
- **Results interrogation** – verifying sensor readings,
- **Milestone notation** – augmenting with supportive context annotations,
- **System maturation** – developing and expertly tuning core processes, and
- **Lifespan innovation** – critical reflection and feedback integration.

This will set the essential stage upon which the real-world outcomes eventually and, naturally, will be judged according to impact and transformative potential.

Returning to Figure 3, we call attention to the main layers, which compose the DQA team levels shaping the aims and objectives that become established targets for the new insight or capability (*center*). This team comprises a protocol and technical teams (*light gray*), project management (*dark gray*), and stakeholder/technology transition (*black*) subgroups.

The Protocol Team. A well-defined question and set of specific hypotheses promote focused data acquisition and more effective troubleshooting. Conversely, ambiguous or loosely stated questions and hypotheses can lead to poor choices of measures, inappropriately defined data acquisition and conditioning parameters, and ultimately, wasted resources, under-powered outcomes, unanswered questions, and limited capabilities.

The Technology Team. While the protocol team is closest to the intent and conceptual forces driving the exercise, the team that is closest to the data and processes that generate it is the technology team, who handle the tasks of calibrating and mounting sensors, implementing and refining data transfer and storage protocols, and conducting analytic checks, ongoing data verification, and successive results validations. Because the technology team is the first line of data generation, acquisition, and handling, their diligence in applying consistent methods and documenting anomalies directly affects whether the data will sustain its quality through subsequent layers of analysis. Moreover, their centrality to DQA also defines their need for clearly articulated, precise, and deeply considered requirements and feedback from the protocol team SMEs and internal leadership.

Project Management. Around this core, internal leadership provides overall direction, decision policies, personnel and fiscal resources, and managerial oversight that either reinforces, refines, or refuses the protocol team's questions and the technical team's software. They decide on budgets, staff allocations, and may heavily affect endpoint evaluation standards, demonstration schedules, and effectively shape how the activities and agenda unfold. For instance, a manager who emphasizes and models thorough, traceable documentation, cultivation of an environment for open communication, and promotion of shared, interactive accountability, can ensure that every iteration of data capture is properly contextualized, supporting continuous improvement and collaboration through the annual and ongoing research lifecycle. Timely, precise, and accurate guidance from leadership also influences adaptability: the team's agility in responding to new insights or shifting mission needs depends on the time and support that they must draw from to enact adaptive responses.

Stakeholders and Technology Transition. At the outer ring, critical stakeholders, like STRL senior leaders, external DoD officials, continental and multinational partners that may include vendors, leaders of industry, university collaborators, and, of course, the end-users—the Soldiers—set overarching demands, compliance rules, performance objectives, and strategic guidance. Their decisions affect what gets measured, expectations about the outcomes of the investment, and how results will be successively communicated and integrated with future decisions or policy. New interoperability standards or integration of a new assessment modality, for instance, might compel the technology team to update data protocols mid-stream, which also then may lead to a cascade of development activity (and the associated unplanned expenses). This outer environment also *dictates the appetite, pace, and direction of technology transition*, a defining function of STRLs. Even if external stakeholders never directly handle or observe raw data, they rely on complete, accurate, and well-documented results to integrate outputs into larger operational contexts and objectives. If stakeholders shift mission parameters late in the process or introduce incompatible requirements, the risk of rework or loss of quality and value detriment increases dramatically; internal teams, *at minimum*, must maintain awareness, adaptability, and compliance with these influences.

By viewing the research question as a living entity influenced by these nested layers, which include the protocol and technology teams, the internal leadership, and the broader stakeholders, we see how DQA is fundamentally a collective undertaking. Responsibility propagates through functionally distinct, hierarchical, and interactive layers. Accountability, therefore, may often become challenging for those further removed from the data endpoint, thus resulting in less inherent connection with the consequences associated with decisions occurring outside the local influence of their team. Alternatively, when each layer respects the evolving needs of a shared objective and upholds a consistent standard for integrity and value curation, STRLs may *systematically* usher innovation from theoretical concepts to robust, field-ready capabilities. This synergy not only secures value in the near term but also reinforces the Army's innovation pipeline, ensuring that data remains actionable, impactful, and poised for transition into real-world solutions.

3. Strategy, Execution, and Tactics for Value Preservation

This section presents a high-level **implementation blueprint** for establishing DQA through a honed and comprehensive process that integrates limited expert resources with advanced data management and software development. Goals do not serve specific *product(s)* but rather are realized in the regular execution of a balanced and adaptive *process* that moves annually from objectives to outcome(s). Earlier sections explained why DQA underpins readiness, innovation, and cost efficiency; here, we emphasize practical means to achieve quality outcomes. A key premise is that *no single role or phase can ensure quality*. Instead, consistency and integrity across the process is essential.

The processes described in this section build on our previous descriptions of team focus (value) and structure (layered ecosystem) to manifest *repeatable methods* that enable the protocol and technology teams to work with more asynchronous leadership structures and broader stakeholders to produce fully aligned and transformative outcomes. Herein, discussion outlines ways to delineate responsibilities and manage accountabilities at each DQA stage, from initial test design to execution, analysis, reporting, and transition. Importantly, clarified tasks and responsibilities enable robust value growth and management through collective accountability; such is *fundamental* for producing datasets that are trustworthy enough to drive strategic decisions.

By the end of Section 3, readers may have a clearer perspective on how a well-planned and mindfully executed DQA structure enables research teams to assuredly transform raw data into the unique value demanded by their respective missions. While fundamentals such as consistent planning, preparation, calibration, transparent metadata, and routine validation remain constant, the specific tools and schedules may change based on experimental goals or external constraints. The subsections that follow will explore these operational building blocks at a higher level, leaving detailed tool descriptions for Section 4.

3.1 Strategic Planning for Continuous Data Quality

A practical way to structure DQA efforts is to follow a long cycle divided into four quarters that map onto annual budgetary and accounting rhythms, which are of course controlled by the U.S. Congress for STRLs—as with other Federal agencies.

Each quarter has a general focus that is appropriate to that time of year and the associated regular activities as dictated by the annual cycle, as summarized in Figure 4. Although real-world deadlines and resource constraints may and often do drift, this quarterly framework provides a reliable **battle rhythm** that serves as a pacing driver for all essential steps, ranging from hypothesis and metrics definition to final reporting. By progressing the team through the successive tasks of gap analysis, implementation, evaluation, and reporting, forward momentum may be maintained without risk of losing key data or insights between major milestones.

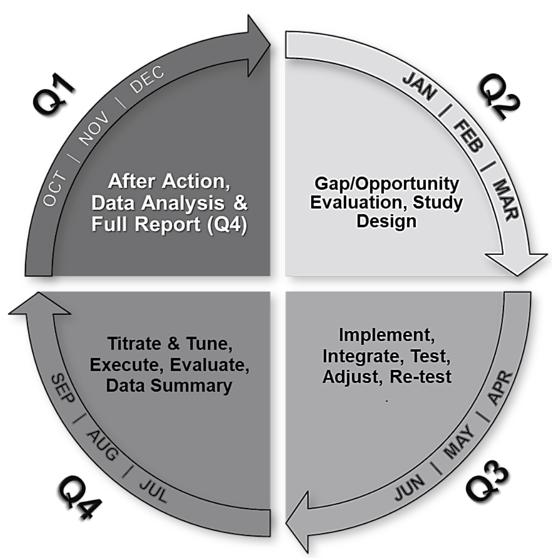

Figure 4. A depiction of the lifecycle of a typical demonstration experiment.

For the sake of discussion, we contextualize the DQA process within the fiscal year used by the U.S. Federal Government, which begins on 01 October and ends 30 September in the next calendar year. Further, for the present discussion, we also consider an annual pace where efforts throughout the year progress toward and culminate in a large, very resource-intensive, and highly visible “demonstration event" (DE), which are often taken as critical opportunities to record uniquely valuable datasets. With this context, **Q1** (Oct–Dec) often concentrates on after-action activities relating to the DE from the previous fiscal year, which will include rapid conduct of prioritized data analyses to address anticipated programmatic deliverables as well as finalizing comprehensive reports that encapsulate the essential programmatic knowledge and technical products. **Q2** (Jan–Mar) then becomes a time of attentional shift, devoted to identifying gaps from the previous year or new opportunities arising from the present technical plan, building and refining study designs, and performing high-level planning for securing of Soldier participants and, of course, undergoing institutional review for the protection of participants. **Q3** (Apr–June) emphasizes an activity shift toward heavy hands-on implementation, testing, and iterative adjustments, along with periodic "sandbox excursions" for technology evaluation based on partial results or pilot feedback. Finally, **Q4** (July–Sep) brings it all together by "titrating and tuning" execution details, conducting the actual DE, evaluating outcomes, and summarizing data that will feed into the next year. Through this, each quarter builds directly on the last, ensuring DQA remains a continuous, spiraling growth cycle rather than a discrete and isolated production process.

3.2 Execution Managed within Directed Action Teams (DATs)

The execution of large-scale, complex experiments requires a structured yet adaptive approach to managing DQA at every stage of the research lifecycle. Given the interdependencies among study design, implementation, validation, and analysis, a rigid, linear process is untenable at STRL scales. A major challenge involves the need to coordinate the dynamic demand for different expertise at different times when the S&Es already have complex task loads. The battle rhythm (see Figure 4) helps. Hypothesis refinement and protocol development occur early, while validation and verification efforts typically ramp up only once sensors are calibrated and pilot data collection is underway. Ultimately, a refined method of coordination is necessary, as DE efforts are typically as follows.

- **Large and multidisciplinary**. Such large-scale experiments and demonstrations are rarely confined to a single discipline, team, or facility. Instead, these efforts demand coordinated expertise across multiple domains, requiring engineers, data analysts, software developers, cognitive scientists, human factors specialists, and operational stakeholders to work cooperatively and efficiently. Unlike a conventional laboratory study, where a small group can oversee the entire lifecycle, DEs are large, distributed, and often involve multiple interdependent phases operating simultaneously. Of course, each of these interdependent phases involve teams and not individuals, leading to a single research effort that may engage two to three dozen personnel, each with uniquely specialized skills and expertise.

- **Supported by in-demand professional SMEs**. Further complicating this situation is the reality that most of the SME personnel working on a DE are rarely fully devoted to that project. SMEs are valued for their ability to contribute across multiple complex STRL efforts—meaning that their time must be carefully prioritized to ensure contributions are maximally effective without creating unnecessary bottlenecks or inefficiencies in the overall workflow. Managing this balance is critical, because an unstructured approach can result in overextended personnel, delays in execution, and gaps in DQA that compromise the integrity of results.
- **Best when self-organized and task driven**. Traditional top-down management structures struggle to keep pace with the dynamic requirements of STRL-scale experimentation. Large, rigid teams attempting to manage a DE often result in bottlenecks, inefficient workflows, and misalignment between available personnel and project needs. Instead, a more modular, targeted approach is required that deploys the right expertise at the right time, entrusting details to the chosen experts, facilitating their deployment without unnecessary effort duplication or expansion.

Achieving a generalized approach involves decomposing DQA, as a whole effort scoped across the year, into a set of smaller, yet still general categories, required tasks. Our solution for this ultimately reduces to a concept we describe as deployable **DATs**, including

- Protocol evaluation and hypothesis clarification
- Measure definition and implementation
- Data validation and integrity verification
- Context annotation
- Runtime quality protection

3.2.1 Phased Deployment and Coordination of DATs

By assigning the *timing criticality* of specific SME attention to tasks *predominantly* when and for as long as they are most organically demanded, and arranging the DATs in fixed sequence with adaptive and overlapping temporal boundaries, regular intercommunication, and effective coordination, a team can ensure that every stage benefits from dedicated, high-caliber expertise and skills, and further that the overall effort progresses normally without excessively drawing on the time and efforts of the people who bear the essential talents. Secondarily, this offers a significant opportunity to build *functional trust* between and among the experts as well as with the DQA team leadership. When promises of limited demand and guarded space for work are respected, overall team integrity grows and often may be rewarded when emergent demands arise unexpectedly during the DE.

DATs operate best as both *interdependent* and *independent* teams. Their interdependence is realized as a planned set of interdictions (checkpoints) among DATs as required for knowledge and product exchanges supporting ongoing efforts. Such checkpoints are coordinated to progress the overall development and execution of the DE as a team in a manner that is organic to any typical developmental process. Figure 5 illustrates, qualitatively, how this might appear for a generalized set of expectations. Despite the clear interdependence just described, these teams are otherwise free to independently select their methods, develop their own internal task structures, group processes, work rhythms, and checkpoints as they deem fit for their assigned tasks.

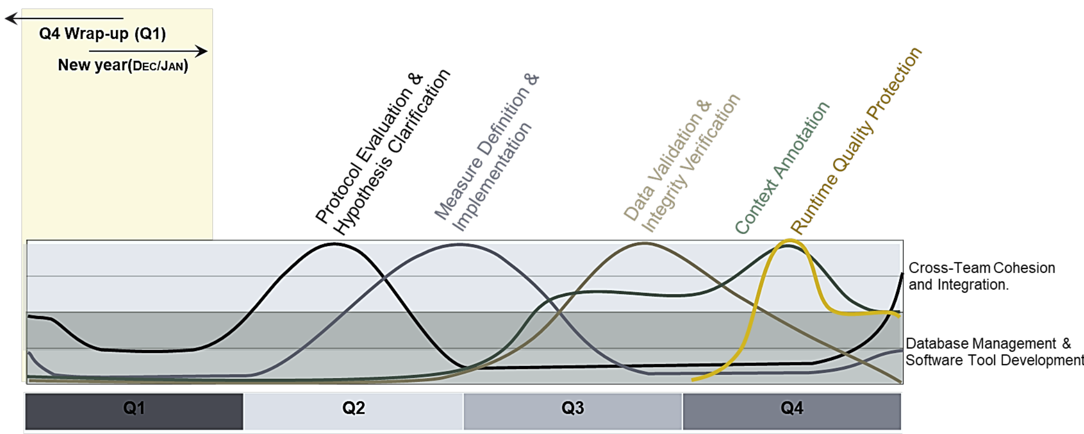

Figure 5. Overlapping phases for DAT deployment through the experiment and demonstration lifecycle. The quarterly structure of the battle rhythm is represented in the labels below the figure, and individual lines indicate relative engagement of each DAT, and shaded background indicates persistent efforts throughout the DE lifecycle.

Moreover, to focus on the checkpoints, it is critical for the DQA team to sustain progress through frequent coordination of **mission essential information** (MEI), whether in the form of synchronous meetings or in time-stamped, threaded, and dedicated (though asynchronous) message boards. Regular checkpoint points should be structured to facilitate communication of priority concerns such as:

- **Status.** Reports of status should include some standard items, such as progress on critical deadlines in DATS, and critical checks on *team*, *timeline*, and *tolerances.*
- **Team** refers to the resources, health, and ongoing workload of each DAT and its constituent SMEs.

- **Timeline** involves management of expectations and negotiation of necessary agenda and schedule modifications as the need arises.
- **Tolerance** is the margin of overall success, such as needed versus available resources to conduct the work, proximity of initial outcomes to targeted results projections, and impending complications or conflicts relative to current team and timeline states.

Who is on point? This should prioritize, always, the urgent and clear sharing of the most critical circumstances, needs, and outcomes related to the ongoing work that is currently engaged, with a DAT that is designated "on point" and working toward achieving the DE quarterly and overall objectives. Frequently, planned efforts encounter unanticipated challenges in various forms. More often, individual DATs can make reasonable predictions about when they expect to reach specific deliverable milestones in the development process, and they are further adept at noting complexities that may require adaptation and coordination in setting up incremental pilots and system tests; thus, their inputs are critical for establishing reasonable initial timelines and coordination expectations.

What is on deck? Among the most critical threats to effective DQA is the handling of transitions in tasking, as well as how the team manages and/or responds to *planned* and *unplanned* informational and product handoffs for successive action or development in waiting. Thus, the team *must be especially sure to effectively manage the transfer and sharing* of products and information that will facilitate the work of the next DAT, focusing on advanced preparation of methods and tools to support specific tasks, technical needs, and team objectives. Critically, design, measurement, and timeline changes need to be decided *in coordination with* any DAT whose successive work may be disrupted; otherwise, significant hidden risks may be incurred. Top-down timelines that were developed independent of the specific experts who will be enlisted are rarely detailed enough for the actual work that must be completed. More specific requirements and timelines need to be generated within DATs, otherwise efforts risk too much time for simple tasks that were assumed to be complex and, likewise, too little time to complete complex tasks that were not understood to be complex. Unfortunately, the consequences of these circumstances tend not to be revealed until times when they are most costly, something that is often avoidable with planful—and even formal—specification and analysis of requirements prior to initiating work (Hobbs et al. 2021).

What will be done, when, and by whom? The greatest weapon that any team has to deploy for protection and assurance of value preservation at the heart of DQA is that of **communication clarity**, which involves precisely stating the

- Exact **tasks** to be completed,

- Exact expected **outcomes**,
- Associated **next-step considerations,**
- Anticipated **consequences** of lack of completion,
- Any **expected difficulties** and resource requests, and
- The current "owner" in the overall **accountability chain**.

These bulleted items may also be recognized as the commonly used SMART acronym, where any given deliverable or goal should be **S**pecific enough that everyone knows when it is or is not achieved, **M**easurable as an achievement in some *objective* manner, **A**ctionable with the current owners' resources, time, and support, **R**elevant to endpoint outcomes and deliverables, and **T**ime-bounded by a specific deadline. Conveniently, regular tracking of the above items within a *structured, and ideally version-controlled, document repository* as the codex of MEI may provide an advantageous resource for later reports that must be generated to describe DE outcomes for stakeholders.

As should be evident, this phase-driven model ensures that each DAT is empowered to draw from a well of specialized STRL expertise, but only what is needed, and only at the right time, reducing inefficiencies and aligning efforts across teams. It also provides an opportunity to document each task or deliverable "owner" to create a **traceable accountability chain**. This structured engagement model further enables scaling resources dynamically, minimizing redundancy, and maintaining high standards of DQA for value preservation, without overtaxing key personnel or creating gaps in execution. Importantly, this also enables better planning, increased team cohesion, supports greater adaptability to unexpected events, and carves a clearer path for transitioning results into actionable and *verifiable* outcomes.

3.2.2 DATs and the Structure of DQA Tactics

Each DAT is assigned a focused problem space aligned to a specific segment of the DQA lifecycle, enabling specialized, time-bounded effort around recurring challenges such as hypothesis clarification, measure definition, data validation, and operational execution. These teams are structured to operate with a high degree of domain fluency while remaining carefully integrated into the overall planning rhythm. What follows is an overview of how such teams are constructed, the types of constraints they are expected to manage, and how their work can shape the structure, reliability, and interpretability of complex experimental data.

Protocol Evaluation and Hypothesis Clarification. One DAT is tasked with ensuring the logical soundness and operational clarity of the research protocol. This team approaches the protocol as a dense, interdisciplinary artifact—often exceeding 70 pages—that must be dissected for coherence across roles, tasks, metrics, and hypotheses. The work begins with close reading and mapping of the protocol structure to expose ambiguities or contradictions, followed by deliberate engagement with the designated "champions" for each research question or hypothesis. These highly focused, interview-based sessions serve to refine research intent, align proposed measures to theory, and eliminate vagueness in outcome expectations. Rather than functioning solely as a check on documentation, this team helps shape the thinking of senior designers and technical leaders by reinforcing the value of fully considered questions, with or without well-specified hypotheses, on the downstream implications for execution outcomes and their interpretation. In many ways, this early-stage clarity becomes the anchor point around which all subsequent DQA efforts are built.

Measure Definition and Implementation. Where the previous team focuses on logical framing, this team takes up the technical and theoretical work of ensuring that what is being measured can, in fact, meaningfully represent the technical constructs and real-world events of interest. This includes ensuring high content validity of the measures and verifying that the resulting data will be statistically reliable and reproducible. Team members often include statisticians, data scientists, and domain specialists who review prior literature, conduct simulations, and iterate on whiteboard models to understand the properties of different quantitative estimators. Importantly, measure development must consider the constraints of the experimental environment, the resolution and fidelity of available sensors, and the eventual interpretability of the data to stakeholders. Implementation may involve developing code modules, data schemas, or runtime scripts that can be embedded directly into collection platforms or postprocessing pipelines. The goal is not simply to generate metrics, but to do so in a way that supports inference, clarity, and value for end users.

Data Validation and Integrity Verification. This team takes as its starting point the assumption that even well-designed data streams are susceptible to breakdowns, inconsistencies, or interpretive ambiguity if not proactively vetted. Here, domain expertise in programming and statistical diagnostics intersects with the kind of judgment required to anticipate how quality issues may manifest. Members of this team define and implement verification tests that will later be used to confirm the integrity and consistency of incoming data, including the generation of test cases or scheduling and administering pilot runs to establish reference expectations. Questions such as *What does a reasonable value look like*? or *How would we detect drift or corruption in real time*? and *What are the natural ranges and distributions of the variables to be computed?* are actively modeled and explored. Some validation strategies may also call for access to sample data from the target context; therefore, this team also is responsible for determining the need, frequency, and nature of pilot tests to be conducted during development. In cases where automated checks are not feasible, and pilot assessments are not feasible, the team designs and rehearses *human-in-the-loop protocols* for runtime verification and shares these insights with the Context Annotation and Runtime Quality Protection teams for use during implementation. This anticipatory stance not only reduces the risk of post-hoc data loss, but also ensures that once data acquisition begins, there is shared confidence that each metric is operating within its intended design space.

Context Annotation. For demonstrations that involve complex behaviors over extended periods—particularly those operating at squad, platoon, or company scale—there exists a need to impose temporal structure and contextual annotation on otherwise raw multimodal data. This DAT comprises trained observers who insert real-time, timestamped labels into the evolving record of each run. These annotations enable subsequent alignment of system logs, behavioral data, and survey responses, and often provide the only usable reference for high-level task segmentation, actor identification, or error event classification. In environments where automated tracking remains imperfect, human annotators offer a bridge between raw telemetry and analyst-ready data frames. The work is intensive, and the cognitive burden is high; thus, team members are supported with purpose-built interfaces, streamlined input modalities, advance training on common annotation processes and observation procedures, and operational briefings that ensure alignment with scenario intent and variable definitions.

Runtime Quality Protection. Operational integrity during live experimentation is often where theoretical assumptions meet real-world limitations and human idiosyncrasies. This DAT is structured around the specific, task-critical points of failure that arise during transitional or high-complexity phases of execution, such as during experiment configuration changes between trials under different conditions. These team members are usually embedded on site, as opposed to working remotely, and are charged with the direct oversight of data-handling touchpoints. These include confirming that sensor streams are initiated correctly, that end-of-run procedures are followed in full, and that validation scripts are triggered as planned. Just as critically, this team sees to the human coordination required across the demonstration team, ensuring that roles are understood, resources are available, and that the Soldier participants are professionally and reliably supported. This is not simply logistics support but is instead focused and mindful DQA at the points of highest volatility. In distributed teams where multiple organizations must synchronize in real time, this becomes central to maintain consistency and safeguard against various controllable errors as well as systemic (often unconscious) drift in protocols and data quality.

Database Management and Tool Development (DMTD). Rather than being formulated as a DAT, DMTD is an *activity* that is also a nexus for all DATs to coordinate. It therefore differs from the others in both scope and continuity. Whereas the DATs are time-bounded and tied to specific phases of the lifecycle, this activity operates continuously, with a core of dedicated personnel (often contracted) who focus their entire role on sustaining it. As shown in Figure 5, DMTD is inherently crosscutting—providing an objective coordination point on which all other DATs may rely. Its primary responsibilities include maintaining schemas, identifiers, and repositories that anchor the work of protocol design, measure definition, validation, and experiment execution. Because every DAT produces or consumes digital artifacts, whether annotations, hypotheses, metrics, or runtime records, this team integrates those flows, enforces version control, and preserves rigorous documentation and audit trails. In so doing, DMTD creates the connective layer that links intent to implementation, ensuring harmonization across efforts, preventing fragmentation as responsibilities shift, and providing the structural continuity necessary for the overall DQA process to function effectively.

Cross-Team Cohesion and Integration. While each team described above operates in a defined space, the overall model assumes intentional overlap and information flow between them. This DQA structure works the best when team members are cross assigned to multiple DATs where possible and appropriate, allowing insight from hypothesis refinement to inform metric definition, or validation strategies to evolve in concert with experiment planning. This structural redundancy supports resilience too; that is, if a schedule shifts or a decision bottleneck emerges, the team has distributed insights rather than siloed their knowledge. More importantly, it reinforces DQA as a *whole-of-effort function* rather than a post-hoc audit task. The alignment across technical, operational, and human-centric dimensions of a complex experiment depends not only on the excellence of each team, but also on the quality of their interactions and cooperation.

Together, these DATs illustrate how discrete technical and organizational roles can be structured into a coherent framework for DQA. Each contributes to a specific function, yet their real value emerges in the ways they interlock and depend upon one another. Having outlined their purposes and interdependencies, we now turn to the practical question of how to establish, resource, and sustain such teams in an operational setting—moving from the conceptual logic of Section 3 to the lived practice that defines Section 4.

4. Task-Driven Tactics for DQA and Army Value Management

Building on the DAT framework outlined previously, here we detail how DQA is continuously interwoven through an entire complicated experimental lifecycle. Thus, this section turns from architectural theory to process manifestation, describing how practical decisions and institutional knowledge converge to form a repeatable, evolvable, and scientifically rigorous DE pipeline.

4.1 Continuous DQA Integration Throughout the DE Lifecycle

STRL experimentation rarely follows a discrete and linear path but is more often cyclical and designed for iterative *capability* refinement and is sometimes envisioned as a *spiral trajectory* over a system or technology concept lifecycle. This means that similar phases of development are revisited, but at increasingly complex levels of integration and understanding, reflecting a commitment to continuous improvement and a deepening understanding of the system under investigation. Running thematically through our team structure and workflows is the core understanding that *DQA is ongoing*—that it is fundamentally an iterative process that must, itself, be maintained with thoughtful practice and goal-directed intention. Certainly, the quarterly breakdown described earlier drives all essential steps, and it further serves as a crucial framework for dynamic resource management and milestone tracking. While this structured approach is especially valuable for resource-intensive experiments and demonstrations, as well as for the inherent challenges of coordinating multidisciplinary teams, it is critical to understand that the quarterly battle rhythm operates *in concert with* the truly continuous nature of DQA.

The annual cycle, along with internal team and task dynamics, establishes synchronization points that specify key *coordination milestones* like hypothesis review, pilot testing, and tiered data-verification points (across organic compute-store cycles). However, fully extracting and curating value requires more granular, ongoing activities and regular synchronization checks at team and interpersonal, as well as systemically defined, checkpoints, reflecting an essential need for strategic oversight focused on broader team alignment and resource management. Fast and tactical responsiveness is needed for team-level awareness, monitoring, and identification of process anomalies or errors that are both critical and urgent.

The DAT structure that we deployed (see Figure 5), with its adaptive and overlapping temporal boundaries, has been key to achieving balance, allowing specialized expertise to be applied at the right time without overtaxing personnel. Importantly, this approach acknowledges the generative nature of experimentation; that is, while guided by structure, there is room for exploration and adaptation within defined boundaries. This necessitates a balance between rigorous planning and flexible adjustment to respond to emergent issues, ensuring that the DQA process supports, rather than hinders, the dynamic evolution of the capabilities under development.

Figure 6 diagrams the overarching workflow within which DQA is embedded—a process that emerged from a more than a decade-long series of Soldier–system evaluations of intelligent technologies for improved situational awareness and decision-making while embedded in simulated or live vehicle-based navigation and mobility tasks (Metcalfe et al. 2010; Cosenzo et al. 2011; Manteuffel et al. 2011). This diagram depicts a sequence of stages, from initial definition of **stakeholder requirements**—defined by Army innovation needs—through to generation of final outputs such as **reports** and **publications**, with critical feedback loops connecting each essential step. A defined capability need is what feeds the **rapid ideation** among program and organizational leadership, which serves to shift focus from the general or notional capability onto **concrete operational concepts** (**CONOPS**) that will provide, ultimately, for a clear **capability demonstration**. This demonstration concept then focuses on the nature and structure of the experimental scenario to be developed and implemented in successive steps.

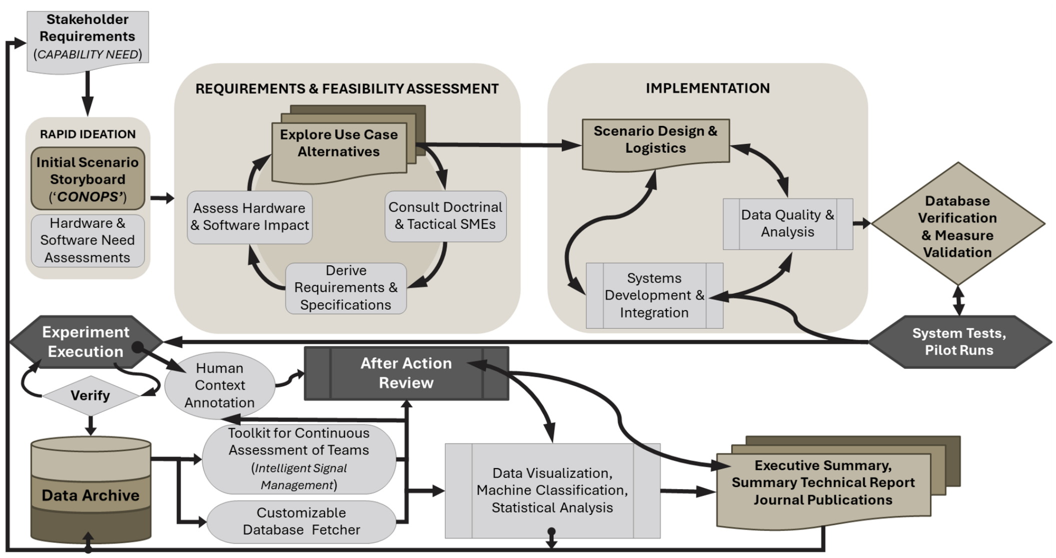

Figure 6. Annualized process for development, execution, and analysis of a mission-based STRL demonstration experiment as deployed by ARL and its collaborating teams for sustained, operationally relevant tests of Soldier–system performance. See text for descriptions.

When CONOPS are settled and agreed, use-case alternatives are generated and examined for issues with feasibility and implementation complexity. Working with appropriate operational communities of interest (e.g., Army operational SMEs, Army Concept Development Integration Directorates, “Concepter Communities,” capability managers, industry and academic experts), along with diligent research and consultation of appropriate doctrinal and tactical materials (e.g., Army doctrine publications available online[*]), the DE team begins defining detailed procedural, logistical, and technical requirements and specifications. As these requirements are formed, a collaborative discourse must occur between protocol and technical teams that will involve a long and involved process of evaluating and addressing emergent hardware and software impacts and their likely effects on performance at runtime. As use cases are proposed and start to take shape, those working on implementation focus on the technical aspects of scenario design and the associated logistical implications; here is where *deep coordination and regular communication must be established between the leadership, protocol, and technical teams* such that important conceptual/logistical changes are aligned to prevent workflow disruptions and to guard against partially informed decisions that may be detrimental to the success of the DE.

Ideally, effort timelines are managed to allow for a sufficiently thorough set of system tests with advance pilot testing targeted at addressing specific developmental decision points and to confirm resolutions of *previously observed and documented* issues, ideally referenceable in an appropriate knowledge base or version-controlled document archive. Indeed, *most* experienced researchers regularly point to pilot testing as both the most important for system robustness, and yet, also *the most commonly given short shrift*; that is—piloting is the phase most at risk of insufficient time for complete bug “shake out,” thus strongly emphasizing the need for runtime quality verification to catch errors due to such potential oversights during earlier phases of development. However much pilot and system evaluation is done, eventually schedules determine that the date of experiment execution will arrive, with the expectation of smooth running of scenarios, intact tools to enable online and post-run data verification, assured routing both to the structured data archive and to After-Action Review (AAR) platforms that utilize data visualization, machine classification, and statistical analysis to generate initial reports and enable development of accurate outcome documents.

By design, this process is purposely iterative, allowing continual integration of new and improved methods and their assessment through *constructive feedback loops* focused on both a) preserving the effective system elements and b) changing those elements needing further adjustment, followed by more testing and validation. Finally, the data archive, supported by tools for continuous assessment, online

[*] https://armypubs.army.mil/ProductMaps/PubForm/ADP.aspx

context annotation, and customizable data fetching, serves as a central repository, for which demonstrable trustworthiness is established by both real-time monitoring and detailed post-experiment analyses (all of which provide for transparent process reports for quality monitoring and improvement), for all curated value from the DE to be shared with stakeholders.

4.2 Lifecycle Tools and Technologies for DQA

Implementing robust DQA within a complex STRL demonstration or experiment requires a coordinated suite of tools and technologies, deployed strategically throughout the experiment lifecycle. Though specific applications, software, and standards will vary across technical and operational domains, we offer the following discussion of tool sets to support readers who may be considering how to establish their own iterative processes of refining hypotheses, defining measurable outcomes, collecting reliable data, and ultimately, extracting meaningful insights. The following outlines key tool categories that best served our work in the ARL INFORMS laboratory.

4.2.1 Protocol Development and Refinement Tools

An effective and complex Soldier-centered DE requires that the leadership team is empowered to maintain a clear and traceable pathway regarding how desired capabilities and CONOPS were translated into variables to observe and hypotheses to test, and finally how these defined the final executable protocol. This begins with establishing a well-structured and collaboratively written protocol, typically scoped by program objectives, directed by project leads, and often written for use in obtaining *required approvals by appropriate authorities* (such as Institutional Review Boards [IRBs], Human Use Councils, and other ethical review panels). The kinds of tools that best support the objectives here include those that enable *iterative development*, *collaborative authoring with version and change tracking*, and *rigorous, multilevel review*. Key tool categories include:

- **Collaborative Document Editing Platforms**: In modern enterprise, these are supported by cloud-based platforms that are configured to enable real-time co-authoring, inline discourse with hierarchical thread adjudication, and finally, accessible version history. Such features are essential for distributed teams working across asynchronously to address sometimes fast-evolving technical demands.
- **Protocol Review and Markup Tools**: Structured feedback systems allow stakeholders to flag ambiguities, suggest revisions, and enforce completeness across required protocol elements, which are especially

important features for assuring alignment with any reviewer or approval body requirements.

- **Hypothesis Formulation and Management Tools**: These help track refinements to hypotheses and aims from initial definition of research into clearly refined and testable statements. They do not require sophisticated tools, but clear structure and traceability. Modern spreadsheet and database tools provide reliable support here, with the benefit of broad familiarity and general usability.
- **Statistical Consultation (or Analysis Tools)**: Access to statistical experts and platforms helps ensure that hypotheses are designed with appropriate *power* (when required), *testability*, and *inferential clarity*. Commonly used frameworks for modern science and computing frequently include R, Python, and other, more specialized frameworks (e.g., TensorFlow for machine learning applications).

Within the DE lifecycle, the Protocol Evaluation and Hypothesis Clarification Team plays a central role in translating the IRB-approved protocol narrative into a practical and testable experimental design. The IRB protocol itself serves as a comprehensive ethical and methodological declaration, covering objectives, procedures, risk mitigations, and subject protections. However, to move from concept to operationalization, this narrative must be decomposed into an internal coordination artifact that is a role usually played by a structured protocol coordination workbook in common spreadsheet software.

For Soldier-In-The-Loop experiments, this workbook may effectively break down the protocol into an easily navigable and intuitive hierarchy of *roles, tasks,* and *metrics*. *Roles* identify all actor categories in the system. Roles may not only include participants, but also experimenters, mission controllers, system operators, and intentional opponent/friendly force or player elements. *Tasks* specify the expected behaviors or responsibilities of each role, which may include system interactions, decision-making events, or observer functions. *Metrics* then define the measurable indicators for each task, including what will be observed, logged, or analyzed to determine whether the task was completed as intended and to what effect. By mapping these elements across all hypotheses, the coordination workbook becomes a living bridge between protocol design, scenario execution, and eventual data capture.

Once this framework is established, the protocol team conducts a series of targeted clarification interviews with each hypothesis champion (e.g., project leaders or other stakeholders who advocate for specific hypothesis or questions to be addressed by the DE). These champions, who should originally be identified during planning to articulate and defend specific research aims, must be engaged in focused dialogues with experiment design SMEs to confirm that their hypotheses have been accurately represented in the role–task–metric mappings in the workbook. Further discourse may then focus on feasibility under anticipated scenario constraints. Where gaps or ambiguities emerge, champions may propose (or take guidance on) appropriate clarifications or adjustments, which are then reflected in updates to the protocol narrative along with (tracked) refinements in the coordination workbook. This iterative engagement ensures that aims are grounded, testable, and appropriately embedded in the scenario.

4.2.2 Measurement Definition and Implementation Tools

Following the establishment of a coherent and testable protocol, the next critical step is translating experimental goals into *valid*, *observable in-context*, and *operationally robust and reliable* measures. This phase involves defining each metric's operational meaning, determining its data source, and ensuring that instrumentation and collection methods are valid and feasible within the constraints of the DE. Tools supporting this work must bridge theoretical **constructs**[*], operational behavior, and the realities of data collection under complex, real-world conditions. Key tools and systems include:

- **Metrics Management Repositories**: Structured spreadsheets or databases for curating the full set of experiment metrics. Each entry typically includes the metric name, definition, unit of measurement, data type (e.g., binary, continuous, and ordinal), expected value ranges, collection modality (e.g., sensor, observer, and survey), and mapping to associated roles and tasks. This repository is often maintained in coordination with the protocol coordination workbook described previously.
- **Data Schema Design and Validation Tools**: Lightweight schema design tools or structured templates help ensure internal consistency across datasets, reduce ambiguity in field coding, and promote compatibility across analysis workflows. These data schemas often mirror or extend existing standards used in STRL instrumentation or Army-relevant data systems, when available, and more importantly are maintained in alignment with transition partners as well as applicable standards (industry, domain).
- **Sensor Calibration and Validation Processes**: Where data capture involves sensors (e.g., GPS, eye tracking, physiological monitors, vehicle telematics, system state logs, and human inputs), it is essential to validate accuracy and sensitivity under expected conditions. Standardized

[*] Typically defined as general traits, states, or capabilities, and ideally have been validated through repeated use in observation and experimentation. Examples include *force*, *intensity*, *awareness*, *workload*, *trust,* and *quality*, to name just a few.

calibration protocols and documentation procedures, even if only locally applicable, will ensure that sensor-derived data streams are comparable across trials and systems.

- **Operational Integration Tools**: Platforms or frameworks that align collected metrics with operational events (e.g., mission phases, scenario triggers, or injects). These tools help link temporal data from disparate sources, allowing post-hoc interpretation of metrics in meaningful context.

In practice, the process of defining and implementing experiment metrics is deeply intertwined with the role–task–metric logic described in the prior section. Once key roles and tasks are agreed upon, candidate metrics are proposed in collaboration with hypothesis champions, scenario designers, and instrumentation specialists. For each proposed metric, the team evaluates whether the necessary conditions exist to capture it reliably: *Is the required behavior observable? Can it be recorded without disrupting the scenario? Is the acquisition technology robust under expected environmental conditions?* This collaborative process often results in multiple iterations of the metrics list, with adjustments made to balance scientific rigor with field feasibility. For example, a cognitive workload metric might be conceptually ideal but practically unmeasurable in austere outdoor settings, prompting a substitution with a behavioral proxy that may be regularly indexed inline and compared with the more traditional measure captured between trials or conditions. Similarly, some desired system interaction metrics may—or will frequently—require custom logging or observer coding, which must be planned in advance.

Ultimately, the finalized metric definitions are documented not only for data collection, but also for downstream use in data integration, quality checking, and analysis. These metric definitions form a bridge between the protocol's experimental logic and the execution team's operational constraints and thus may ensure that the data collected are not only *relevant*, but also *valid* and *reproducible* and are therefore *trustworthy*. By institutionalizing such a tool-supported metric development process, STRL teams increase the transparency and rigor of their DEs, enabling stronger inferences and broader applicability of findings.

4.2.3 Context Annotation and Runtime Verification Tools

Reliable data interpretation in field-based DEs hinges not just on whether data were captured, but on the ability to interpret those data in the correct operational context. This requires a system of tools that enable real-time annotation of ongoing activities, structured capture of experimental deviations, and verification that protocol-aligned activities occurred as expected. Tools in this category facilitate the semantic enrichment of raw data, enabling *traceability* and *situation awareness* across execution teams. Key tools and systems include:

- **Structured Annotation Interfaces**: Custom-built interfaces that allow designated observers or experiment controllers to easily and accurately apply consistent schema-governed annotations during live execution. These interfaces typically support structured tagging of events with information like actor identity, intended target or recipient, task quality, scenario or mission execution state, and may also enable other protocol-relevant attributes, in real time or shortly after occurrence.
- **Observer Cue Cards and Job Aids**: Printed or digital aids designed to reduce cognitive load for annotators by surfacing key scenario tasks, expected behaviors, and known failure points. These tools help standardize observer expectations and improve annotation consistency across team members and runs.
- **Runtime Scenario Dashboards**: Centralized displays that show the real-time status of scenario elements, including actor progression through tasks, synchronization events, or deviation flags. These dashboards help scenario leads verify execution integrity and detect errors or drift as they happen.
- **Deviation Logging Systems**: Lightweight structured forms or checklists used by scenario controllers to record any deviation from the planned protocol (e.g., skipped tasks, delayed coordination, and instrumentation failure). These logs maintain linkages to original scenario intent and support later interpretation during analysis.
- **Annotation Schema and Integration Frameworks**: Backend tools for maintaining annotation schemas and linking live annotations with logged data, metrics, and post-run summaries. This supports later aggregation, filtering, and alignment with hypotheses and evaluation metrics.

In the INFORMS laboratory DEs, structured context annotation played a pivotal role in ensuring that scenario execution remained aligned with the research intent. Trained observers used custom annotation tools (Figure 7) to tag key scenario moments, such as task start/stop events, mission pauses, scenario anomalies, moments of especially high- or low-team cohesion, and specific expressions of technology-based frustration. These tags were inserted directly in the dataset with appropriate timestamps to enable later alignment with other key variables and events. Each annotation was predefined and agreed to by the *annotation team*, those who performed this role during the DE, ensuring consistency in how events were recorded across runs and observers. Concurrently, a runtime dashboard displayed task status and synchronization checkpoints, allowing the scenario coordination team to catch issues such as role absence, sequence violations, or missed triggers. When deviations occurred, structured forms allowed the lead controller to document the nature, timing, and suspected impact of each incident in a standardized way. These tools collectively supported high-fidelity reconstruction of each run and gave downstream analysts the context required to determine whether experimental conditions were sufficiently preserved to support valid inference.

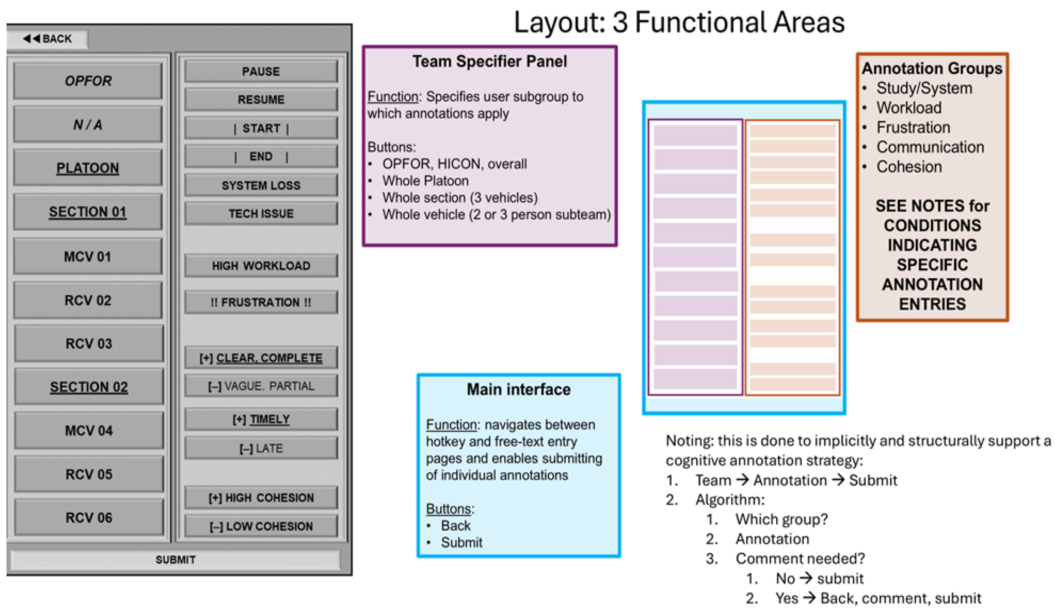

Figure 7. Example of a structured annotation interface used during an ARL INFORMS Soldier touchpoint.

Observers were provided with role-specific cue cards (Figure 8) detailing the expected behaviors for various actor categories (e.g., participants, mission leads, and oppositional forces), along with a separate structure form to enable taking process notes (a self-monitoring DQA technique) to allow noting task-specific events and common sources of confusion or breakdown, for possible later use in data cleaning and correction. These served as live job aids and pre-briefing guides, enabling a shared mental model across the observer team.

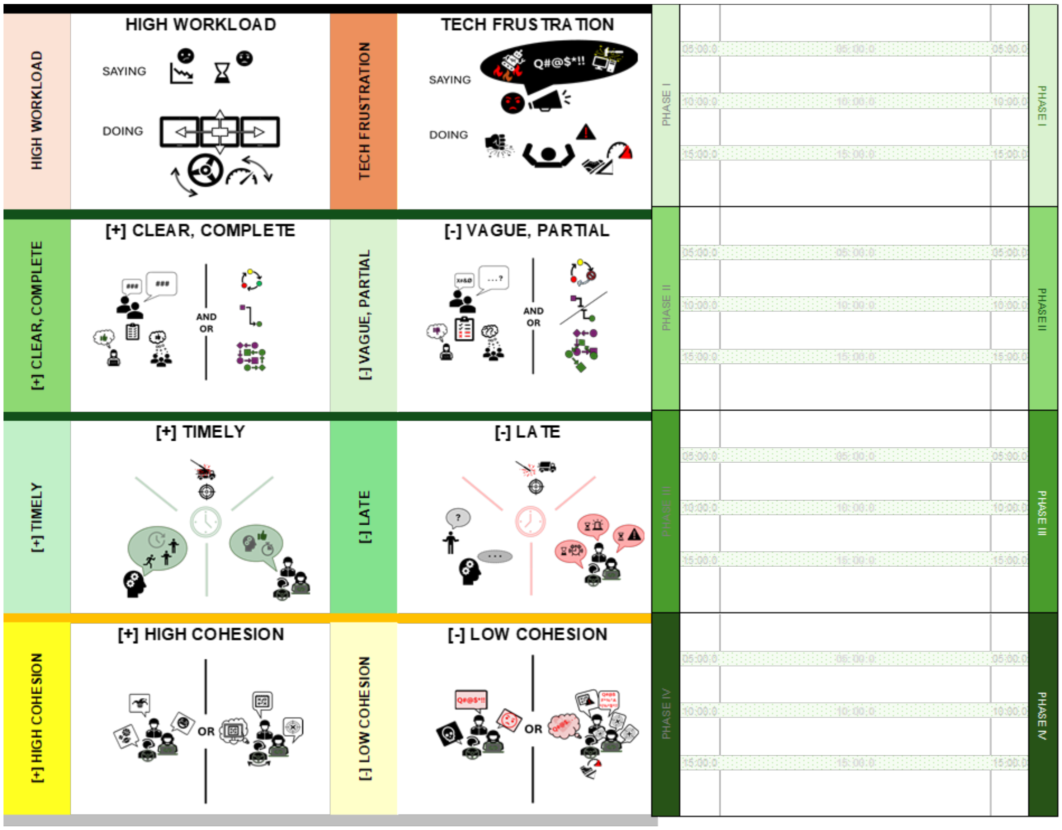

Figure 8. Visual cue-card (left) and pre-structured observer-annotator process notes (right) corresponding to the annotator interface shown in Figure 7.

4.2.4 Post-Run Analysis and AAR Tools

Once experimental data were collected and annotated, the focus of DQA shifted to post-run analysis and structured review. This stage held a focus on whether the data generated were sufficient to support the research hypotheses, questions, or aims and, moreover, whether they delivered valued insights usable by both scientific and operational stakeholders. Herein the goal is twofold: 1) to identify measurable outcomes with statistical rigor, and 2) to enable teams and leaders to reflect constructively on how scenario elements, technologies, and human interactions contributed to success or failure. Key tools and systems include:

- **Statistical Analysis Environments**: Tools such as R, Python (with *pandas*, *scipy*, and *statsmodels* libraries), and domain-specific tools (e.g., SAS, SPSS) that enable rigorous examination of the relationships between experimental conditions and observed metrics. These platforms are essential for hypothesis testing, correlational analysis, and evaluation of treatment effects under varying operational contexts.
- **Data Visualization Frameworks**: Libraries like *ggplot2*, *seaborn*, *matplotlib*, and dashboarding tools such as *Tableau*, which allow analysts

and stakeholders to explore the data visually. Alternatively, custom dashboards may be more appropriate to develop in specific contexts. These tools are critical for surfacing patterns, contextualizing anomalies, and communicating findings to both technical and nontechnical collaborators and stakeholders alike.

- **Session Trace Tools and Timeline Reconstruction**: Custom scripts or integrated platforms that reconstruct the timeline of a single experimental run from multimodal data (e.g., system logs, observer annotations, and participant telemetry). These tools are particularly useful for verifying task sequences, identifying misalignments between scenario design and real-world behavior, and building narratives around what actually occurred.
- **AAR Integration Tools**: Software that aligns observational data, quantitative outcomes, and qualitative feedback (e.g., participant interviews, and facilitator notes) into a common timeline or analysis frame. This supports structured AARs that are data-informed rather than anecdotal, enabling lessons learned to be grounded in operational evidence.
- **Behavioral Analysis and Team Assessment Toolkits**: Purpose-built toolchains like ARL's Toolkit for the Continuous Assessment of Teams (T-CAT; Figure 9) or the Human–Autonomy Team Trust Toolkit (HAT^3; Neubauer et al. 2022a, 2022b), which offer structured constructs for evaluating coordination, communication, trust, and performance across hybrid teams. These toolkits integrate well with structured role–task–metric models and extend the ability to assess socio-technical system effectiveness over time.

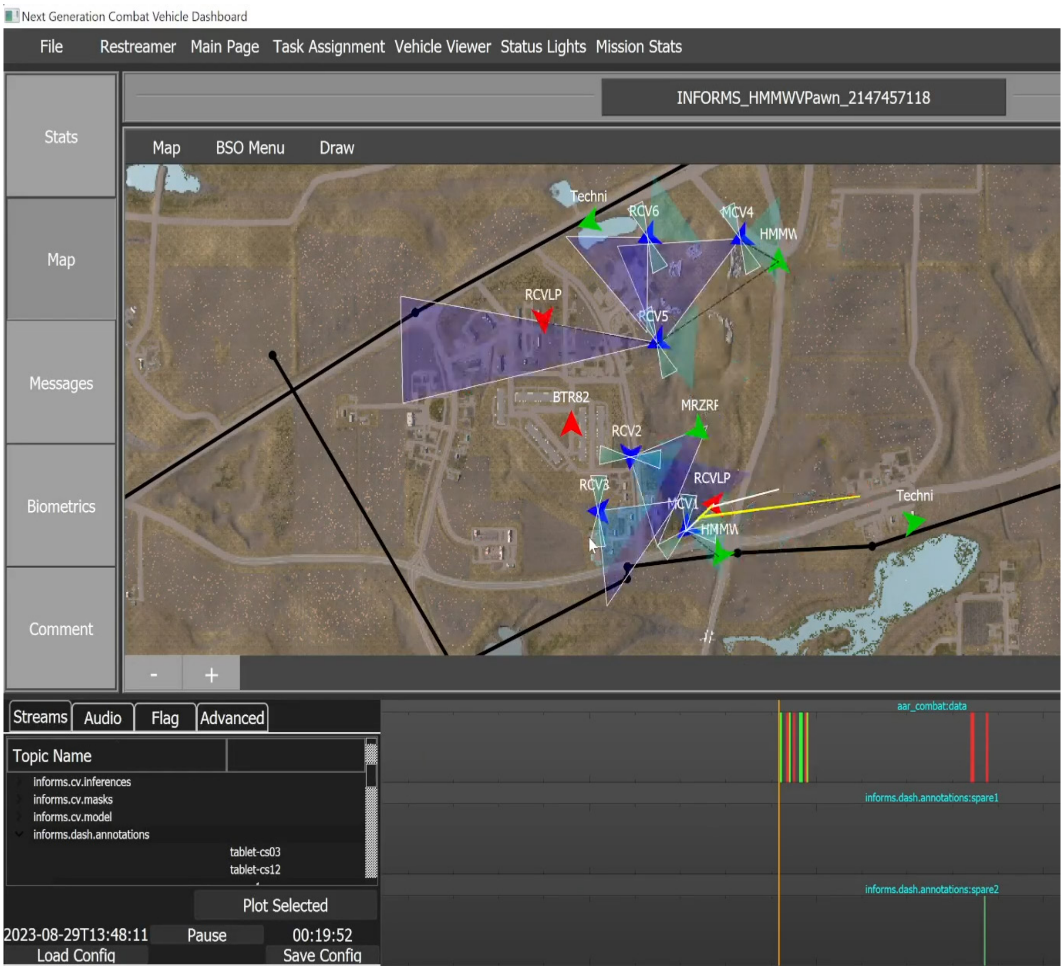

Figure 9. Example of a mission observation/playback tool integrated into ARL's T-CAT as deployed in the INFORMS laboratory.

In the ARL INFORMS DEs, post-run analyses and AAR preparations drew upon a multilayered data ecosystem built during repeated phases of the lifecycle over repeated DE executions. Raw logs, structured annotations, scenario deviation reports, and survey instruments were harmonized into a consolidated analytical dataset, anchored to the role–task–metric schema developed during protocol refinement. Analysts employed Python-based pipelines for data wrangling and exploratory statistics, while statistical inference was performed using both frequentist and Bayesian models depending on metric sensitivity and sample size.

A key innovation was the integration of annotation streams with runtime metadata to produce per-run trace narratives. These timelines enabled experiment leads and scenario controllers to revisit specific interactions and decision points with clarity, identifying which behaviors were planned versus emergent. For example, timeline reconstructions were used to trace when coordination failures occurred and what preceding events or environmental conditions contributed to them—insights that would have been inaccessible through metrics alone. AAR sessions incorporated these traces into SME-facilitated discussions, supported by visualization overlays and behavioral assessment modules from the T-CAT suite. Soldier participants and observers could jointly review performance against defined expectations and operational intent, fostering shared understanding of system behavior and surfacing design or execution adjustments needed for future runs. This data-grounded AAR process served not only to evaluate experimental outcomes, but to iteratively improve the DE process itself.

Together, these four categories of tools—protocol development, measurement implementation, runtime annotation, and post-run analysis—formed a practical scaffold for embedding DQA principles across the full lifecycle of an advanced technology demonstration and experimentation. Each was conceived not as a one-size-fits-all solution or as a finished product, but as modular and evolving component tools that teams may apply and tailor to their technical constraints, experimental objectives, and organizational structures. By documenting not just the tools but the rationale for their integration, we aimed here to provide a reusable and extensible framework that other STRLs can adapt to their own contexts. With the technical groundwork established, Section 5 synthesizes key lessons and strategic considerations for sustaining high data quality over time, particularly as teams scale their operations and seek to institutionalize these practices.

5. Sustaining Data Quality: Key Insights and Recommendations

Consistent execution of a complex DE within an STRL environment requires more than just robust tools, it demands a commitment to established best practices and a willingness to learn from past experiences. Recurring themes from logs of past project meetings within the team that used this DQA process will highlight several key areas critical to our overall success.

Proactive protocol management and revision control is *paramount*. The team consistently emphasized the importance of early and frequent review cycles for the protocol document. This includes designating clear ownership and accountability for revisions, as exemplified by detailed marking of required changes including subsequent task assignments to team members to maintain the accountability trace. Equally critical is the prompt resolution of technical issues impacting document access, for instance, repeated efforts to re-upload and ensure accessibility of the protocol despite sometimes uncertain network or collaborative component vendor stability. A tight deadline, such as the IRB submission date, necessitates a focused and rapid turnaround for addressing reviewer feedback and adaptability when otherwise reliable tools become disruptions.

Effective communication and collaboration were thematically essential to maintenance of the DQA *process*, recalling that process stability led to improved efficiency and reliability in arriving at outcomes. Clear task assignment and tracking, facilitated by tools like accountability sheets for hypothesis development, are essential for maintaining momentum. Encouraging cross-team input and integrating diverse expertise, particularly in areas like robotics and human–robot interaction, enriches the quality of the research. Open communication channels are vital for quickly resolving issues, as evidenced by discussions surrounding data-related concerns and logistical challenges.

Prioritizing traceable documentation and accessibility is foundational. Maintaining comprehensive documentation of hypotheses, metrics, and engagement personnel lists ensures transparency and facilitates knowledge transfer. Equally important is providing reliable access to critical documents, requiring proactive troubleshooting of document management systems and swift resolution of technical issues. The team's emphasis on preparing amendments to address changes in questionnaires and hypotheses underscores the need for a flexible and responsive documentation process.

Iterative tool improvement and adoption must be held and nurtured as an ongoing, living, and continuously evolving process. DQA only improves as processes improve. The focus on enhancing data annotation standards, driven by a desire for clearer communication and improved interface usability, demonstrates a commitment to refining tools based on user feedback. This includes exploring academic benchmarks for annotation quality and implementing user-friendly interface improvements. Furthermore, a "guided self-organized fashion" is crucial for effectively implementing these robust quality checks while adhering to defined timelines and accountabilities.

These lessons learned, consistently reinforced throughout the project lifecycle, provide a valuable framework for future STRL DEs. By prioritizing proactive planning, open communication, meticulous documentation, and continuous improvement, research teams can maximize the quality and impact of their work.

6. References

Bronfenbrenner U. The ecology of human development: experiments by nature and design. Harvard University Press; 1979.

Bronfenbrenner U. Ecological systems theory. In: Vasta R, editor. Annals of child development. JAI Press; 1989. Vol. 6. p. 187–251.

Cosenzo K et al. Impact of 360° sensor information on vehicle commander performance. Army Research Laboratory (US); 2011 Mar. Report No.: ARL-TR-5471.

DAMA International. The DAMA guide to the data management body of knowledge (DAMA-DMBOK2R). 2nd ed. Technics Publications, LLC; 2024.

Floyd D. Spence National Defense Authorization Act for fiscal year 2001. H.R.4205. Pub L No.: 106–398, 114 Stat 849; 2000 Apr 6.

Good B. Army's Stryker double v-hull is a resounding success. PEO Ground Combat Systems (US); 2012 Nov 30. https://www.army.mil/article/92154/armys_stryker_double_v_hull_is_a_resounding_success

Hobbs KL, Davis J, Wagner L, Feron E. Formal specification and analysis of spacecraft collision avoidance run time assurance requirements. 2021 IEEE Aerospace Conference (50100); 2021 Mar 6–13; virtual. IEEE; c2021.

Laney DB. Data juice: 101 stories of how organizations are squeezing value from available data assets. O'Reilly Media; 2022.

Manteuffel C et al. A methodology for the assessment of 360° local area awareness displays. In: Guiell JJ, Bernier KL, Thomas JT, Desjardins DD, editors. Proceedings of the Display Technologies and Applications for Defense, Security, and Avionics V; and Enhanced and Synthetic Vision 2011. SPIE Defense, Security, and Sensing; 2011 Apr 25–29; Orlando, FL. Vol. 8042. SPIE; c2011. p. 126–139.

Metcalfe JS et al. Experimentation and evaluation of threat detection and local area awareness using advanced computational technologies in a simulated military environment. In: Gerhart GR, Gage DW, Shoemaker CM, editors. Unmanned Systems Technology XII. SPIE Defense, Security, and Sensing; 2010 Apr 5–9; Orlando, FL. Vol. 7692. SPIE; c2010. p. 72–83.

Neubauer C et al. Human–autonomy teaming (HAT^3) executive summary. DEVCOM Army Research Laboratory (US); 2022a Dec. Report No.: ARL-TR-9622.

Neubauer C et al. Human–autonomy teaming (HAT^3) trust toolkit software development documentation and user guide. DEVCOM Army Research Laboratory (US); 2022b Dec. Report No.: ARL-TR-9623.

List of Symbols, Abbreviations, and Acronyms

AAR	After-Action Review
AI	Artificial Intelligence
ARL	Army Research Laboratory
CONOPS	Concrete Operational Concepts
DAT	Directed Action Team
DE	Demonstration Event
DEVCOM	U.S. Army Combat Capabilities Development Command
DMTD	Database Management and Tool Development
DoD	Department of Defense
DOTMLPF-P	Doctrine, Organization, Training, Materiel, Leadership, Education, Personnel, Facilities, and Policy
DQA	Data Quality Assurance
FY	Fiscal Year
GPS	Global Positioning System
HAT	Human–Autonomy Teaming
HAT^3	Human–Autonomy Team Trust Toolkit
INFORMS	Information for Mixed Squads
IoT	Internet of Things
IRB	Institutional Review Board
LLM	Large Language Model
MCV	Manned Control Vehicle
MEI	Mission Essential Information
OPFOR	Red Team
RDTE	Research, Development, Test, and Evaluation
S&Es	Scientists and Engineers
S&T	Science and Technology

SMART	Specific, Measurable, Actionable, Relevant, and Time-bounded
SME	Subject Matter Expert
STRL	Science and Technology Reinvention Laboratory
T-CAT	Toolkit for the Continuous Assessment of Teams

1 (PDF)	DEFENSE TECHNICAL INFORMATION CTR DTIC OCA
1 (PDF)	DEVCOM ARL FCDD RLB CI TECH LIB