Tackling the Human Performance Data Problem: A Case for Standardization

Alexxa Bessey, Luke Waggenspack, Brian Schreiber
Aptima, Inc.
Woburn, MA
abessey@aptima.com, lwaggenspack@aptima.com, bschreiber@aptima.com

Winston Bennett, Jr.
Air Force Research Laboratory
Wright Patterson, OH
winston.bennett@us.af.mil

ABSTRACT

Over the past few decades, simulator use has increased greatly, due in part to its cost-efficiency and ability to provide training experiences that would be impractical or unsafe to conduct otherwise (e.g., emergency procedures). This increase in simulator use has coincided with an explosion in "big data," more specifically, human performance data that are collected from a large number of learners (n), measured variables (v), and measurements per unit time (t) (Adjerid & Kelley, 2018). However, as the resulting corpus of human performance data expands, it becomes increasingly more difficult to mine for trends, resulting in a large pool of recorded data that is not immediately useable without extensive workarounds, manpower, or software algorithms. For example, consider the use case of simulated Air Force engagements. At any single Air Force training facility, there could be simulator records from hundreds of training scenarios per year with a variety of different characteristics (e.g., offensive counter-air maneuvers, defensive counter-air maneuvers, two-ships, four-ships, etc.). However, certain limitations of the data, such as unstandardized start and stop times of the engagements, hinder the ability to easily mine the data for historical norms, proficiency, or other human performance outcomes. As a result, the ability to interpret or draw conclusions from the data is much more limited, despite the robust pool of data. In this paper, we present the findings from a multi-year research and development effort that focuses on extracting meaningful human performance metrics from a "data lake" of roughly 3,500 data recordings that represent 10,000 training scenarios over the course of more than 15 years. We present best practices and lessons learned for parsing the data lake contents so that readers can better understand the implications of data limitations and how to address them in their own work.

ABOUT THE AUTHORS

Alexxa Bessey is Scientist in the Training, Learning, and Readiness Division at Aptima, Inc. She has over 5 years of experience conducting research within military environments, including her time spent as an operational research psychologist in the field. Ms. Bessey offers an expertise in assessment, unobtrusive measurements, training, and teams. In addition, she is well-versed in data collection and sampling methodologies in military settings, to include both subjective and objective approaches. At Aptima, Ms. Bessey is involved in several projects including the assessment of proficiency-based training, the evaluation of simulator concurrency, the examination of unobtrusive measures in teams, and the development of a sleep application for elite Soldiers. As part of her work at Aptima, in addition to her doctoral work, Ms. Bessey's research examines validating unobtrusive data measurements. Ms. Bessey has a master's degree in both Clinical Psychological Science and Industrial-Organizational psychology and is currently completing her PhD in Industrial-Organizational psychology from Clemson University.

Luke Waggenspack is an Associate Scientist at Aptima, Inc. With a PhD in Educational Psychology from Texas Tech University, Dr. Waggenspack is trained in state-of-the-art methodology and data analysis techniques, with a specialty for missing data analysis. He has experience consulting for and conducting research across a wide variety of realms of scientific inquiry. He has made personal contributions to both the education and methodological literature in the form of publications and software. His current work focuses on topics such as learning in modern technological environments, the creation and testing of new statistical analyses, and complex data and analytical structures.

Brian Schreiber is a Principal Scientist with Aptima. He holds a master's in science from the University of Illinois at Champaign-Urbana and has been performing research and development within the military domain for over 27 years. He has authored or co-authored over 60 papers, journal articles, tech reports, and book chapters.

Winston Bennett Jr. received his Ph.D. in Industrial Organizational Psychology from Texas A&M University in 1995. He is currently the Readiness Product Line Lead for the Airman Systems Directorate located at Wright Patterson AFB Ohio. He has been involved in NATO-related research activities for over 20 years. He has also been involved in I/ITSEC committee and program work for a number of years as well. He is spearheading the Combat Air Forces migration to proficiency-based training and is conducting research related to the integration of live and virtual training and performance environments to improve mission readiness and job proficiency. He leads research that has developed methods to monitor and routinely assess individual and team performance across live and virtual environments and evaluating game-based approaches for training, work design, and job restructuring. He maintains an active presence in the international research and practice community through his work on various professional committees and his contributions in professional journals and forums including I/ITSEC. His involvement with the larger psychological communities of interest ensures that communication amongst international military, industry and academic researchers remains consistent and of the highest quality.

Tackling the Human Performance Data Problem: A Case for Standardization

Alexxa Bessey, Luke Waggenspack, Brian Schreiber
Aptima, Inc.
Woburn, MA
abessey@aptima.com, lwaggenspack@aptima.com, bschreiber@aptima.com

Winston Bennett, Jr.
Air Force Research Laboratory
Wright Patterson, OH
winston.bennett@us.af.mil

INTRODUCTION

As a result of advances in technology, training devices used by the United States Air Force (USAF), to include simulators, have resulted in an influx of recorded data. As the emphasis on the value of training and human performance data has increased, organizations such as the USAF have prioritized the collection of a large corpus of human performance data in order to inform important outcomes such as proficiency-based training. While such large datasets are beneficial, there are several notable challenges to properly managing and analyzing the data. Unfortunately, such issues are often the result of small decisions, oversights, or a failure to communicate data analysis priorities in the initial stages of data collection and storage, all of which can result in adverse and costly consequences when trying to analyze the data. As a result, the following paper discusses important considerations when collecting human performance data as well as a case study outlining the analysis of a data lake of over 10,000 training exercises. More specifically, the following paper will highlight how a small feature in the data, the lack of standardized exercise start and stop times, severely minimized the initial utility of the data and lead to the implementation of an extensive algorithmic workaround. Further, the objective of this paper is not only to present a unique case study to demonstrate the importance of thoughtful data collection processes but also to present a challenge to the military community on how to progress human performance data analytic capabilities moving forward.

The Power of Playback: Recording Human Performance Data

As technology capabilities expand, additional opportunities are available to collect and interpret diverse and extensive human performance data. For example, in many domains (e.g., health care, professional sports, the military), the increase in personalized and affordable wearable sensor devices (e.g., Apple watches) has resulted in a corpus of health-related data that can be tracked longitudinally. Additionally, the increased accessibility and decreased cost of storing large amounts of data has further encouraged the creation of large datasets (Fan et al., 2014). Both of which (e.g., more accessibility, decreased costs) will only continue to facilitate additional data collection moving forward. As a result of this trend, there has been a rather drastic prioritization on the consumption of data to inform research through collecting, mining, and analyzing large datasets (Fan et al., 2014).

While the idea of high-dimensional and large datasets is exciting and presents several potential advantages, previous research has also highlighted some of the challenges that come with the exponential expansion of data, including how to manage and interpret large amounts of organizational data, as well as the challenges of relying solely on data-driven approaches to interpret the data (Fan et al., 2014; Orvis et al., 2013). Further, across most domains, there is a general consensus that capturing as much data as possible is a best practice. This notion may fail to address, and therefore simplifies, some of the complexities and challenges that come with analyzing large datasets. More specifically, a lack of forethought on the front end of data collection processes, such as how data is captured and stored, can limit the usability of the data without extensive workarounds when it is later analyzed on the backend. This particular issue highlights the notion that oftentimes it is easier to collect data than it is to quickly find useful and meaningful assertions from the data (Tsai et al., 2015). As a result, while the overall sentiment of collecting large datasets represents a valuable effort, overlooking certain features of the data that limit or prevent analyses can misrepresent the value of datasets without extensive workarounds.

The following paper addresses this issue by describing a case study of a large corpus of human performance data from recorded USAF simulator training sessions. More specifically, how certain features of the data, stored within a large data lake, severely limit the usability of the data without extensive algorithmic workarounds. As it relates to the value

of human performance data, if human performance data is not immediately usable or has several limitations, then the value of that data is also limited. In the case of the USAF and the military at large, human performance data that is restricted presents as an operational disadvantage and limits the ability to accomplish proficiency-based training and other human performance informed outcomes.

CASE STUDY

Background

In recent years, the USAF, like many other military organizations, has become increasingly reliant on the use of simulators to provide training. Simulators are advantageous in that they offer a more budget-friendly approach to training while also minimizing adverse safety outcomes (e.g., crashes). Further, simulators are able to mirror more complex threats and emergency procedures and therefore can provide a heightened level of readiness for operators. As the use of simulators has increased, recordings of simulator training exercises have also increased, resulting in a source of valuable training data. While previously such recordings were used almost exclusively for warfighter debriefs, the push for longitudinal and dynamic training data has resulted in a multitude of tools that can record and play back the exercises (e.g., the Live, Virtual and Constructive Network Control Suite [LNCS]) as well as tools that can extract data from the recordings (e.g., Performance Evaluation Tracking System [PETS]; Schill et al., 2014; Schreiber, 2013). In essence, the USAF has begun to mirror other industries (e.g., professional sports) that have prioritized archiving and analyzing recordings as a way to inform training and other performance outcomes above and beyond debriefs. As the USAF and the military at large seeks to improve data analytic capabilities of the increasing corpus of human performance data, the number of such tools will also likely expand.

Further, advances in data mining approaches have provided additional opportunities to examine training and human performance outcomes. In many cases, such as the case study described below, data from the recordings as well as the raw files of the recordings have been deposited in a data lake for future analysis. However, despite the general sentiment that capturing more data is beneficial to understanding different outcomes, issues within the data collection and storage process can limit the immediate usability of large datasets, suggesting that only considering the collection of large datasets is not sufficient in order to extract meaningful training and human performance outcomes. As a result, the following case study discusses a 15-year research effort that has resulted in a data lake of 3,500 data recordings that represent 10,000 training exercises. The following case study will first discuss an overview of historical issues with the data as well as provide a more granular examination of an issue regarding the lack of standardized exercise start and stop times. Although highly specific, the objective of this examination is not only to speak to the community most affected by this issue, but also to present a use case for others to proactively consider similar issues with human performance data and the consequence of those issues in their own work.

Documented Human Performance Issues

Too numerous to go into detail in the current paper, there are a number of issues that can come into play naturally with previously archived historical data; we briefly mention them here. The first, and most common, is loss of information. Many historical documents about training which was taking place at a given time is separate from the recordings of the data itself and difficult to accurately link the documentation to the appropriate recording. In many cases, the document itself is also missing. As such, any given recording might not contain or may be missing key information about who was flying in it, where it was flown, what type of scenario was flown, weapons loadouts, and the list goes on. Additionally, historical data may be more prone to bugs and errors that are more difficult to troubleshoot given the gaps in time between data collection and analysis. For example, two common, very problematic, examples include the velocity and acceleration vectors necessary for calculating G-load not existing properly in the data. In addition, in some cases, munitions send mistimed detonation and deletion messages resulting in extra, not real munitions to appear in the data. While such issues drastically impact the usability and quality of stored data due to prevalence of missing information or data, the lack of standardized start and stop times within the data provide an example of a significant yet simple problem that has sweeping consequences to the analysis of human performance data. More explicitly, how recording simulator training scenarios were started and stopped (and therefore archived) has drastically impacted the utility of large datasets above and beyond missing data or time-related errors.

The Importance of Start and Stop Times

Although several issues have been noted when analyzing archival data from simulator training exercises, a particularly simple but significant issue is the lack of standardized exercise start and stop times. During recorded simulator training exercises, instructors frequently activate or "start" entities (e.g., platforms), which pushes a protocol data unit (PDU) across the network. In some cases, this happens to all entities (a "global start") and in other cases this happens to just one or a handful of entities (a "start" or "unfreeze"). During the exercise itself, the functionality of this feature is to activate and manipulate different entities for the different training scenarios (relevant to the instructors) and for the PDUs to accurately communicate across the system (relevant to software engineers who create and manage the network). However, the functionality of the start and stop times outside of the exercise doesn't map on the functionality needed for data analysis such that the start and stop times fail to capture the actual periods of performance within the exercise. For data analysis, the functionality of starts and stops represent the period of time to capture and summarize performance data, much akin to tabulating points/statistics according to a sporting game time "clock" starting and stopping at the beginning and end of a game. Quite problematic for analyses then, the start and stop times recorded in the data are represented by an unstandardized smattering of start and stop PDUs, that are either present, not present, or inconsistent (multiple starts, no stops, etc.). The end result is a very large database of scenarios (thousands) in which the scenarios cannot be automatically partitioned correctly to the actual start and stop times. If automatic partitioning cannot occur, then this completely prohibits aggregating the database for analyzing summaries across scenarios and therefore, creates the inability to derive performance outcomes from the data.

To illustrate, consider an NBA basketball game. If you were to record the data from a regulation NBA basketball game, it would last 48-minutes. However, if you "started" recording points 10 minutes early (i.e., capturing baskets "scored" during warm-ups) and "stopped" the data capture 30 seconds early, the outcome of who won or lost could be different (i.e., wrong) than if you correctly started/stopped a data recording for the game. The incorrect start and stop times of the game WILL drastically alter the performance outcomes and assertions made regarding what happened during the game itself would be flawed. Now consider you were to go to Madison Square Garden and continuously recorded data such that you captured multiple games. You would then have an aggregate of all the outcomes, such as points scored across multiple games, instead of outcomes for just one game. In that case, even mediocre players could rack up an impressive +50 points. More specifically, if you were to look at the aggregate data as if it were a single period of performance, or game, your analysis of the game and the players would again be gravely misinformed. In this case, the lack of start and stop times for each relevant period of performance creates an inaccurate aggregate dataset. In both cases, analyzing simple statistics for the season would be wrong; for example, the total wins/losses for each team would be incorrect. Additionally, any given player's performance per game (e.g., shots scored) would also be incorrect. As a result, the ENTIRE database has lost much of its analytical value, due solely to incorrect start and stop times.

Similar to the basketball example above, the lack of a standardized exercise start and stop times within simulator training recordings has resulted in two key problems for analyzing human performance data, both of which have detrimental implications for drawing conclusions regarding proficiency-based training. First, a lack of standardized exercise start and stop times can fail to identify the actual period of performance of the exercise. Meaning, the lack of a standardized start and stop time results in the entirety of the recording being captured, which can include (1) extraneous events within the recording (see Figure 1) or (2) truncate important events (see Figure 2). As a result, any analyses may also include events outside of the period of performance (similar to analyzing the basketball game with the warmup included) or fail to include important events (similar to analyzing the basketball game after failing to record the first quarter). In either case, failing to identify the correct period of performance within the recording can result in an inaccurate analysis of the data such that your outcomes may include data outside of the period of performance and/or may truncate the period of performance, therefore excluding key events. Such a problem is then compounded when you have thousands of scenarios in a data lake.

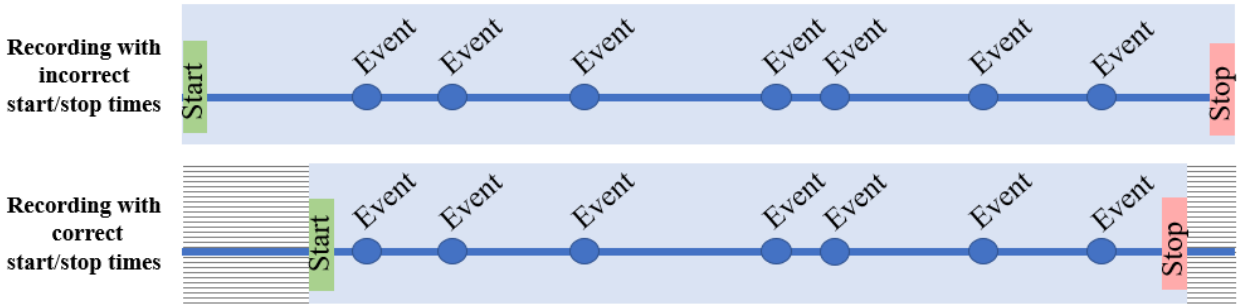

Figure 1. Identifying Exercise Periods of Performance (Extraneous Data)

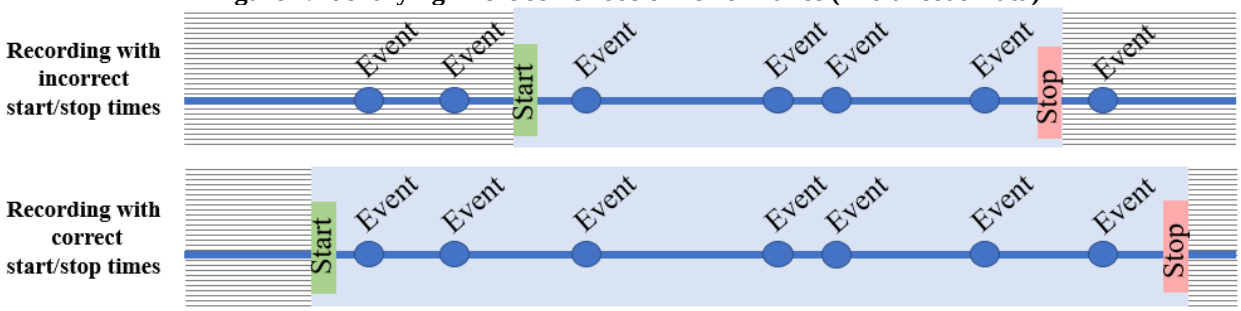

Figure 2. Identifying Exercise Periods of Performance (Truncated Data)

Next, in any single recording, there may be one exercise or several (e.g., an hour-long session may have several scenarios). Since standardized start and stop times of the exercises were not implemented when recording the data, the beginning and end of each individual exercise within the recording is not indicated. This issue is particularly relevant when a recording has multiple exercises. Because the exercises within the recording are not properly indicated, the end result can be the incorrect analysis of human performance data (similar to recording multiple basketball games but analyzing the performance data as if it were just one game). Meaning, the lack of a standardized start and stop time results in the entirety of the recording being analyzed as one exercise instead of multiple exercises (see Figure 3). More specifically, outcomes such as flight time, shots fired, and kills would be drastically different if the recording was being analyzed as one exercise opposed to correctly analyzed as three separate exercises. In this case, the lack of standardized start and stop times of the exercises results in an unreliable period of performance which limits the ability to analyze and glean meaningful insights from the data. Similarly, this issue can further compound with the previously described issue such that not only are multiple exercises within a recording being aggregated, but there also is substantial extraneous and/or truncated data within the recording. The end result is a dataset in which human performance outcomes cannot be confidently extracted without extensive workarounds and data cleaning, both of which can be time consuming and costly.

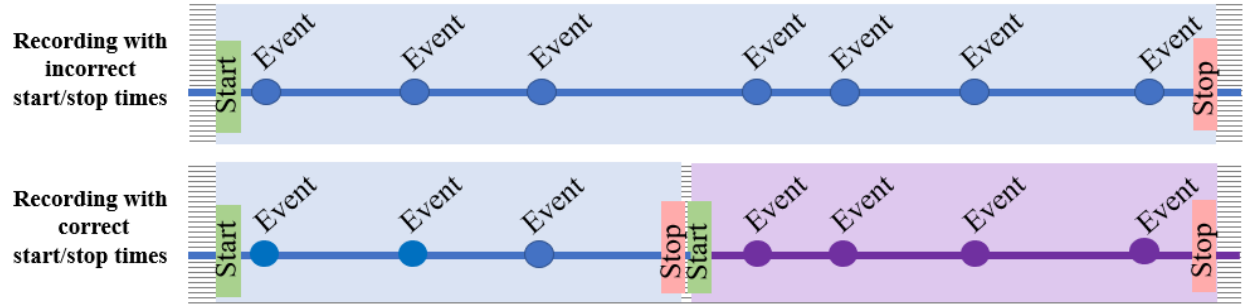

Figure 3. Identifying Exercise Periods of Performance (Multiple Exercises)

In summary, unstandardized exercise start and stop times provide an example of a tangible and simple issue that has massive implications that impact the ability to analyze the data for meaningful human performance outcomes. As a result of a lack of systematic use of start and stop times in simulator training recordings, the stop and start time of each

exercise, and more specifically the desired period of performance within each exercise, was not documented. Since a recording may have several exercises and/or partial exercises (e.g., an exercise was reset) as well as extraneous "dead time" (e.g., set up of simulators and features of the exercise), periods of actual performance are not immediately clear and therefore cannot be used to extract meaningful and accurate training and human performance outcomes. Simply put, the consequence of a lack of a standardized process to record start and stop times has rendered the data limited without extensive workarounds such as software algorithms. In doing so, the data is either (1) not immediately useable and therefore the intended outcomes are unable to be derived or (2) the data is used incorrectly, and the outcomes derived from the dataset are based on incorrect periods of performance and therefore are wrong. In other words, human decisions, and the lack of standardized processes on the front end of data collection prevent or severely undermine the benefits of collecting large datasets without costly consequences.

Start and Stop Solutions

As a result of the lack of standardized start and stop exercise times, additional workarounds had to be implemented in order to properly analyze the data. In our research, we developed an analytical approach (a "start and stop logic"), in order to correctly and automatically identify exercise start and end times (e.g., periods of performance) within a large data lake of archived data. Given the previously described two primary concerns regarding start and stop times, the purpose of the logic was twofold. The first purpose of the logic is to trim the exercises to include only the relevant period of performance from each exercise. Therefore, the logic extracts only the appropriate data, or period of performance, from each exercise for analysis. Second, the logic sought to correctly identify and separate multiple scenarios/exercises within one recording. In doing so, the logic provided the ability to identify the correct periods of performance for analysis. Taken together, the goal of the logic was to identify/partition individual exercises and capture the meaningful periods of performance within such exercises while simultaneously avoiding excluding key performance data.

The initial design of the algorithmic logic was primarily based on creating a weapons range for all entities (e.g., F-35) such that the exercise start time would be triggered when the first set of entities were within a designated range from each other. Likewise, the exercise stop time would be triggered when the last set of entities were outside the designated range. Weapons ranges were identified as the primary priority given the importance of shots as a key feature of the period performance as well as a key performance outcome. The range distances used to inform the logic was created by multiplying the notional weapons range of each individual platform (e.g., the weapons range of a F-16) by a max range factor developed by subject matter experts (SMEs). The max range factor was created in order to identify weapon employment ranges without disclosing actual shot range capabilities. More specifically, the calculated range would create a buffer such that the range was exclusive enough not to trigger too early when entities were extremely far away but inclusive enough to not miss any key events (e.g., shots fired). Depending on the platform, most of the ranges used to inform the logic were between 50-80 miles.

While the logic in its original iteration was fairly successful, additional modifications were added to the logic in order to refine the accuracy of the logic. For example, initially, logic for long range weapons had to be modified to address pre-mature triggering due to too large of a range (e.g., triggers happened immediately). Further, of in more complex distributed training scenarios, entities that were being dragged into place were triggering the logic with inactive entities while being moved. As a result, the logic was modified such that both entities had to be active to trigger the logic.

In order to test the efficacy of the start and stop logic, 25 historical recordings were examined. Each scenario was run on the most recent version of PETS to get a computer-generated start and stop time for each scenario. Separately, a SME watched each scenario giving it a start and a stop time. The start and stop times indicated by the SME were then compared to the start and stop times that were selected by the logic. For the start times, 100% of start times were within acceptable ranges (5 seconds or less difference when compared with the SME), demonstrating high efficacy of the program at generating proper start times. In the case of stop times, however, only 60% were within acceptable ranges. However, many of those outside of the acceptable range were only slightly outside the acceptable range (7-10 seconds). In either case, the logic did not miss any important events taking place, indicating that while the logic may still be capturing some extraneous data (e.g., 7-10 seconds worth), the logic is successful at capturing all necessary events within the period of performance.

Lessons Learned and Next Steps

In the process of analyzing the large corpus of human performance data, several key insights and lessons learned noted. To begin, when developing the algorithm that informed the logic, the initial basis for the development of the logic prioritized shots, as previously mentioned. Although the rationale for this was sensical based on the importance of shots as a key performance outcome, focusing on one outcome failed to consider the implications of other outcomes (e.g., triggering the logic early in the case of long-range weapons). More explicitly, by focusing on one goal or objective instead of the collective goal served initially as a barrier to the development of a more comprehensive logic. Another key lesson learned from this process is the importance of developing a validation plan. At several stages of the development of the logic, pre-planned validity checks were implemented to better understand the progress of the logic. For example, after the initial logic was developed, a multi-disciplinary team (e.g., engineers, scientists, SMEs) observed the logic's ability to implement start and stop times in a series of randomly selected exercises. As a result, the validation process was how the team corrected for errors.

In addition to the lessons learned related to developing the logic, two much broader experiences occurred that provided key insights for the team, both of which have direct implications for other multi-disciplinary teams interpreting human performance data. As previously mentioned, the functionality of the start and stop times differed based on which group was being considered. More explicitly, for the instructors, the purpose of the start and stop times were to manipulate the training exercise. However, in the case of the engineers, the start and stop times were more of a manifestation of PDUs accurately communicating within and across the networks. Further, the purpose of the exercise start and stop times for the scientists was to correctly identify a period of performance in order to analyze the data for performance outcomes. For example, because the ultimate goal of the scientists was to define the correct period of performance, the use of "global" start and stop terminology was initially utilized to refer to the start and stop of the desired period of performance. In contrast, and based on terminology utilized in the Distributed Interactive Simulation (DIS) standards, global start and stop times more accurately reference the type of PDU pushed such that a global start represents a PDU push that enables all entities. As a result, the scientists and engineers had a much different interpretation of global start and stop times which served as a barrier to both communicating and problem solving. Therefore, the functionality of start and stop times within the team had to be accurately and precisely defined in order for team members to correctly interpret both problems and potential solutions for analyzing the data. This particular example highlights the need for multi-disciplinary teams to examine and standardize language in order to ensure that all team members are approaching the problem with the same lens. More explicitly, multi-disciplinary teams should ensure that goals are accurately communicated across the team members and a standardized set of definitions is created for terminology that may be domain or discipline specific.

Similarly, interpretations of problems with the logic needed to be standardized across the team to accurately capture what aspects of the logic needed to be modified. For example, in the initial iteration of the logic, there were still issues with capturing the correct period of performance due to activity of entities pre-exercise triggering the logic too soon. In this case, the logic was triggering a start as the engineers had intended based on the specifications outlined by the SME, however, the goal of capturing the correct period of performance was still not achieved as outlined by the scientists on the team. This differentiation resulted in additional issues with communication as the logic was simultaneously working correctly (for the engineers) and incorrectly (for the scientists) depending on which objective was focused on. While it is pertinent for team members to understand the key objectives for their own tasking, team members should also understand key objectives for other member's tasking in order to wholistically comprehend and solve the problem. This also requires a synthesis of knowledge across multi-disciplinary team members in order to identify partial solutions that fail to address the collective goal. For example, even though the logic was initially functioning, it required the input of the SME who was familiar with simulator exercises for the team to understand that the logic was not capturing the actual periods of performance. In other words, it required each team member to contribute their own understanding of the problem as well as their own expertise to understand that the initial solution was not working. Once the problem was fully outlined for all members of the team, the team was able to work together to create additional guidance for the logic (e.g., requiring both entities to be active) in order to prevent the previously described problem.

Taken together, the lessons learned highlight the need for highly functional multidisciplinary teams when tackling human performance data. Given the need for rigorous research methodologies, insight from an experienced SME, as well as the need for highly technical software engineering skills, having multidisciplinary teams is crucial for success. However, in order for multidisciplinary teams to be successful, it is pertinent that team members are able to communicate and comprehend problems and solutions across the team.

CONCLUSION

In conclusion, while human performance data can be extremely valuable, collecting, storing, partitioning, and managing the data must first be considered on the front end in order to efficiently analyze the data on the back end. More explicitly, collecting data without proper foresight of how certain features of the data may impact analytic capabilities may limit the extensive efforts of data collection and for future use, prevents scaled growth. To that extent, the paper outlines a case study describing the lack of standardized start and stop times when completing simulator training exercises and the costly and time-consuming consequences of the lack of standardized data collection processes. Such issues are costly in that not only do they require time (e.g., years of work) and effort (e.g., extensive labor across multiple disciplines) to fix, but also that the extraction of key performance insights that could be used to inform training are delayed. As a result, the USAF is left with an untapped resource that could and should be used to improve training and therefore, mission performance. Stated another way, the inability to immediately draw certain human performance conclusions from previously recorded data is slowing down a more comprehensive and well-informed approach to proficiency-based training. Which certainly impacts USAF training outcomes, but also may have more consequential effects as training outcomes transitions to mission outcomes.

Identifying correct periods of performance are crucial for analyzing human performance data as incorrect periods of performance can drastically impact insights derived from the data. This can include instances such as the aforementioned case study (e.g., unstandardized start and stop times) as well as other use cases across military and performance-oriented research. As a result, researchers should leverage the previously described case study when outlining their own data collection and analysis methodologies. Further, the developed logic and validation process may have additional implications for archived human performance research, especially in instances where periods of performance may be identifiable based anticipating the occurrence of certain events (e.g., utilizing weapons range to inform capturing shots fired).

As it relates to the USAF and military community at large, more formalized, and standardized processes must be implemented in order to bolster human performance analyses and inform training outcomes, such as proficiency-based training. More explicitly, a great emphasis must be made on converging the priorities of the military at large (e.g., well-versed training data) with the priorities of the operators collecting the data. While operators must first consider how to effectively implement and execute training in real time, there also must be a strong emphasis on data collection methodologies and standard practices in order to facilitate performance analyses. This should be done while considering both the importance of standardized processes as well as the burden of the operator. In conclusion, this case study highlights the important lesson that just because data is being recorded and stored in a large dataset or data lake does not mean that the dataset is immediately useable for many types of analyses.

REFERENCES

Adjerid, I., & Kelley, K. (2018). Big data in psychology: A framework for research advancement. *American Psychologist*, *73*(7), 899–917. https://doi.org/10.1037/amp0000190

Fan, J., Han, F., & Liu, H. (2014). Challenges of Big Data analysis. *National Science Review*, *1*(2), 293–314. https://doi.org/10.1093/nsr/nwt032

Orvis, K., Duchon, A., & DeCostanza, A. (2013). Developing Big Data Based Performance Measures: A Rational Approach. *The Interservice/Industry Training, Simulation & Education Conference (I/ITSEC)*, vol. 2013, Orlando, FL.

Schill, N.P., Rowe, L.J., Gyovai, B.L., Joralmon, D.Q., Schneck, A.J., & Woudstra, D.A. (2014). *Operational alignment in predator training research* [Paper presentation]. The Association for Unmanned Vehicle Systems International, Orlando, FL.

Schreiber, B. T. (2013). Transforming Training: A Perspective on the Need for and Payoffs From Common Standards. *Military Psychology*, *25*(3), 294–307. https://doi.org/10.1037/h0094970

Tsai, C.-W., Lai, C.-F., Chao, H.-C., & Vasilakos, A. V. (2015). Big data analytics: a survey. *Journal of Big Data*, *2*(1). https://doi.org/10.1186/s40537-015-0030-3