Standardizing Human Performance Measurement for Ease of Data Analytics

Beth F. Wheeler Atkinson, Mitchell J. Tindall	John Killilea	Michael Tolland, Courtney Dean
Naval Air Warfare Center Training Systems	Stracon Services	Aptima
Orlando, Florida	Orlando, FL	Woburn, MA
{beth.atkinson, mitchell.tindall}@navy.mil	john.killilea.ctr@navy.mil	{mtolland, cdean}@aptima.com

ABSTRACT

As interest grows for big data analytics within the Department of Defense (DoD), one prime opportunity to leverage existing data sources is performance assessment. Specifically, the use of quantitative performance data for determining skill levels of trainees supports diagnostic feedback, targeted remediation, and identification of opportunities to accelerate or tailor training to student learning progress. The successful implementation of automated, system-based performance measures within DoD training systems for assessment and trend analysis purposes, however, necessitates standardization in implementation to ensure success. Based on on-going efforts, the authors propose two areas for consideration: 1) adoption of standards for hardware and software simulation interoperability, and 2) an approach to measurement definition that is flexible to the military's *crawl-walk-run* approach to training and conducive to trend analysis. Currently, the simulation community lacks a standardized way to represent human performance data requirements that are generalizable, scalable, interoperable, and transparent. Because of this gap in standards, developers are challenged with finding ways to implement technology in environments that lack the right type of data. The first step toward increased consistency would be an industry standard for interoperability. As such, this paper will outline a proposed human performance measurement standard under consideration by the Simulation Interoperability Standards Organization (SISO). This standard provides a framework for defining how a system can utilize available data to determine if trainees achieve desired outcomes based on the mission context. However, because not all facets of human performance measurement can be defined by a standard, researchers and developers must consider other factors during measure implementation. For example, measures may be presented in the form of raw data to inform instructor formulated assessments (e.g., number of kills), or assigned values to automatically classify calculations (e.g., percentage, expert vs. novice). Both of these forms of measurement provide unique data benefits throughout the training lifecycle, but a theoretical approach to defining and implementing performance measurement for trend analysis is required to fully realize those benefits. Therefore, in addition to a proposed interoperability standard, the authors will provide lessons learned and best practices for performance measurement when long-term goals include pursuing big data analytics.

ABOUT THE AUTHORS

Ms. Beth F. Wheeler Atkinson is a Senior Research Psychologist at NAWCTSD and a NAVAIR Associate Fellow. She has led several research and development efforts devoted to investigating capability enhancements for training and operational environments, and has successfully transitioned a post mission reporting and trend analysis tool that leverages automated performance measurement technology. Her research interests include instructional technologies (e.g., performance measurement, post-mission reporting/review), Human Computer Interaction (HCI)/user interface design and analysis, and aviation safety training and operations. She holds an M.A. in Psychology, Applied Experimental Concentration, from the University of West Florida.

Mitch Tindall, PhD, is a Research Psychologist at NAWCTSD in the BATTLE Laboratory. He works in several areas such as HCI, data management and analytics, training systems enhancement and validation, and systems software evaluation. His Ph.D. is in Industrial-Organization (I-O) Psychology from University of Central Florida (UCF).

John P. Killilea is a Research Psychologist supporting the Naval Air Warfare Center Training Systems Division (NAWCTSD) in the Basic & Applied Training & Technologies for Learning & Evaluation (BATTLE) Laboratory. He holds a Masters in Modeling & Simulation, and is currently working on his Ph.D. in the same field at UCF.

Michael Tolland is the technical lead for Aptima's Performance Measurement Engine and the Director of Product Engineering. He leads the development of performance measures for that engine, and integration the PM Engine with numerous simulation environments. Mr. Tolland holds a B.S. in Software Engineering from the Rochester Institute of Technology.

Courtney Dean is a senior scientist at Aptima, Inc. Mr. Dean has extensive experience in survey development, job analysis, and test construction and validation, with a focus on social networking and rating convergence in assessment centers. Mr. Dean holds a M.S. in Applied Psychology from the University of West Florida and a B.S. in Psychology from Fort Hays State University and is a member of the Society for Industrial and Organizational Psychology.

Standardizing Performance Measurement for Ease of Data Analytics

Beth F. Wheeler Atkinson, Mitchell J. Tindall
Naval Air Warfare Center Training Systems
Orlando, Florida
{beth.atkinson, mitchell.tindall}@navy.mil

John Killilea
Stracon Services
Orlando, FL
john.killilea.ctr@navy.mil

Michael Tolland, Courtney Dean
Aptima
Woburn, MA
{mtolland, cdean}@aptima.com

INTRODUCTION

While data analytics or *big data* is no longer a new phenomenon, the Department of Defense (DoD) training community is just beginning to investigate the benefits within the military. There is immense consensus regarding the value of data in informing objective decisions, which can result in organizations collecting more data than they know how to effectively manage (McAfee & Byrnjolfsson, 2012). Thus, the efficient and effective use of large data sets remains a challenge. Major David Blair (2015) of the United States Air Force wrote about this topic in the Navy Quarterly, discussing how rich data sets already exist in military aviation environments but that data is tied exclusively to the aircraft or simulator providing little analytic value on the performance of aviators. The premise of his argument is that relying on these rich data sets in training allows us to move beyond the *check in the box* participation credit for training. Instead, it allows us to embrace a culture that leverages detailed performance data to develop more effective, proficient aircrews. This presumption about the value of big data in performance assessment is consistent with theory and empiricism of academics that have studied it for a long time. Specifically, when feedback provided to trainees is accurate, timely, specific, and directed at the task, it is far more effective in improving subsequent performance (Kluger & DeNisi, 1996). The military's increased reliance on simulation to train the warfighter, coupled with technological advances offering Live, Virtual, and Constructive (LVC) environments provide increasing opportunities for collection of automated system-based performance data. That is, we no longer have to rely solely on the observations of instructors to assess and debrief trainees. While there are benefits from consistency and standardization of assessment, it is important that automated performance measurement does not preclude instructors from assessing the processes that led to successful or unsuccessful outcomes. Rather, efforts to develop automated performance measurement tools should be seen as a supplemental assessment capability. In the following paper, we will discuss an effort to develop a standard for leveraging large data sources from complex training systems to generate Measures of Performance (MOPs). These measures could be used both for developmental purposes and for analyzing larger trends that provide valuable information to decision makers. The effort to standardize available network data to assess performance represents an opportunity to move away from stove-piping and toward modularity in proficiency tracking across integrated training events at individual, crew, team-of-team, and organizational levels of assessment.

HISTORY OF INTEROPERABILITY

Individual and crew-based training are crucial in ensuring warfighter knowledge, skills, and tactics necessary to be effective in operational settings. However, no individual or aircrew operates in a vacuum. Mission success is always dependent on the coordination and communication of multiple entities (e.g., command and control, surface vessels, other aircraft, allies). For this reason, warfighters engage in integrated warfare training events after gaining tactical expertise at the individual and/or platform level to *train as we fight*. Traditionally, integrated warfare events occur in live settings. The costly, dangerous, and logistically challenging nature of organizing and executing these events results in infrequent large-scale, live training opportunities. In response to these limitations, DoD moved to distributed simulation-based events, which require the hardware and software systems in these environments to operate in concert with one another for effective training. Challenges emerge when components developed in a stovepipe and/or by different organizations do not fully consider interoperability—the ability of systems and software to exchange and make use of information to work with other systems or products.[1]

[1] From "Standards Glossary," by the Institute of Electrical and Electronics Engineers, retrieved from https://www.ieee.org/education_careers/education/standards/standards_glossary.html. Copyright 2017 by IEEE.

As networked simulation increased, organizations saw the need to establish standards that facilitate interoperability of systems and increase the quality and efficiency of modeling and simulation. Protocols such as Distributed Interactive Simulation (DIS) and High-Level Architecture (HLA)[2] resulted. Within the Naval Aviation community, the Naval Aviation Simulation Master Plan (NASMP) was established to address interoperability of Fleet Aircrew Simulator Training, providing guidance for distributed simulation within the Navy Continuous Training Environment (NCTE) by outlining trainer requirements to implement HLA to achieve interoperability. Recent years have seen continued advances in achievement of distributed training, providing opportunities that leverage an appropriate mix of cutting edge LVC infrastructure and capabilities (Naval Aviation Enterprise, 2016).

As new concepts, goals, or technologies emerge, organizations such as the Institute of Electrical and Electronic Engineers (IEEE) and the Simulation Interoperability Standards Organization (SISO) seek to facilitate the establishment of standards and products to increase capability and interoperability of emerging and future technology. Similar processes exist within Navy Aviation, with federation changes managed by the NASMP Federation Working Group. Examples of common changes to the federation include objects and interactions that enable the display of new platform capabilities on the simulation network (Wiese, Atkinson, Roberts, Ayers & Ramoutar, 2012). In general, new standards and updates to federations focus on the engineering aspects associated with a shared environment to ensure consistent implementation. For example, IEEE's HLA 1516-2010 provides an overarching framework and rules for a common architecture in simulation (see IEEE Standards Association website), while SISO's Standard for Common Image Generator Interface focuses on a specific aspect of the environment, visual presentations (see SISO website for SISO-STD-013-2014). While this is critical to interoperability goals, there is an additional need to supplement those standards with protocols focused on instructional capabilities. An example of this type of effort, which resulted in an approved SISO standard, would include the Standard for Distributed Debrief Control Architecture (DDCA; see SISO website for SISO-STD-015-2016), which establishes the means to facilitate distributed debrief discussions essential to the learning process through shared mechanisms for DVR-like controls that sync replays. However, standards to provide an infrastructure for handling of Human Performance Measures (HPMs) are still lacking. As a result, developers have difficulty implementing technology in domains that do not possess data or information specific to measuring human performance.

HUMAN PERFORMANCE MEASUREMENT

As the DoD seeks mechanisms to optimize readiness, leadership recognizes the need to "deliver integrated and interoperable warfighting capabilities that produce an immediate and sustainable increase in warfighting effectiveness" (Naval Aviation Enterprise, 2014, p. 12). The justification for this data-driven decision making has been in part due to a need to maintain readiness as live flight hours are reduced (Buss, 2014). By providing analytic tools that enable informed decision making on the mixed use of simulator and LVC infrastructures, the Navy is able to meet the demands of reduced flight hours with additional benefits of increased training opportunities and reallocation of training to live environments where this level of immersion and/or capability is absolutely required (Naval Aviation Enterprise, 2016). Specifically, tools such as HPMs provide the means to assess the structured use of simulation-based training. Further, HPMs facilitate the feedback process to ensure an effective and efficient training event (Stacy, Freeman, Lackey, & Merket, 2004). For example, Astwood, Van Buskirk, Comejo, & Dalton (2008) found a 10-20% improvement in training effectiveness and enhancement of fleet readiness when diagnostic feedback is provided. The typical methods for collection of HPMs include self-report (i.e., directly from trainees), observational (by instructor-observers), or through automated, system-based technologies (Wiese et al., 2012).

Developing self-report or observational performance measures is a well-established process in academia and industry; however, administering, completing, and analyzing such data requires significant manpower (MacMillan, Entin, Morley, & Bennett, 2013; Wiese, Nungesser, Marceau, Puglisi, & Frost, 2007). Additionally, ensuring reliability across observations and self-assessments is an almost futile effort. Recent research geared toward improving inter-rater reliability by incorporating error awareness training (i.e., teaching awareness of errors that negatively affect accuracy) or frame-of-reference training (i.e., teaching consistency in construct and metric

[2] From "Overview of SISO: Who we are and what we do," by the Simulation Interoperability Standards Organization, retrieved from https://www.sisostds.org/AboutSISO/Overview.aspx. Copyright 2017 by Simulation Interoperability Standards Organization - SISO.

definition) still only produced coefficient alphas no greater than .80 (Schleicher, Mayes, Day & Riggio, 2002). While this level of reliability is not excessively low, achieving a similar level of consistency across instructors in a military context with various confounding factors (e.g., lack of co-location, personality, culture, competing priorities) is likely near impossible. Conversely, Automated Performance Measures (APMs) that leverage system data may provide instructors with information that is consistent and optimal for providing detailed diagnostic feedback.

The implementation of APMs requires access to information for measurement computation. Technological advances and increased employment of simulation standards in general provide a mechanism to publish vast amounts of raw data that can be used for APMs (Stacy, Ayers, Freeman, & Haimson, 2006). However, with the rapid proliferation of distributed and simulation-based training, the sheer amount of information available makes identifying the *right* data difficult (Portrey, Keck, & Schreiber, 2006). In order to achieve distributed training, existing standards focus primarily on issues pertaining to interoperability of hardware and software systems to provide a shared environment and experience. To date, standards have lacked consideration of student performance measurement leading to insufficient progress in providing easy access to data that supports determination of training effectiveness. Specifically, this means data related to HPM in simulation environments is often difficult to identify or impossible to find because other federates do not publish data relevant to calculating APMs (Portrey et al., 2006). As a result, systems fielded often lack capabilities to standardize assessment through the implementation of APMs (Atkinson & Killilea, 2015). One means to rectify this is through a standard specific to HPM.

Delaying the establishment of an HPM standard, as more complex training systems come online, is likely to result in immense cost in both time and resources (Wiese et al., 2007). When training effectiveness results are desired, training professionals will need to conduct front end analyses that may involve lengthy and costly processes to customize measures for each individual effort. Defining a universal standard of system-based HPM that is incorporated in system specifications at the inception of development should alleviate the problem of sifting through the large breadth of data available for calculating measures across systems (MacMillan et al., 2013). Additionally, with proper standards, a shift in sole reliance on instructor-observed performance to a system that leverages instructor and system-based performance assessment is possible. As a result, the community will gain the ability to employ data-driven analyses to decipher the subtleties of aircrew proficiency with multiple benefits.

Consistency

When there is inconsistency in the way raters understand constructs over time or across groups (i.e., violation of configural invariance; e.g., Buss & Royce, 1975; Irvine, 1969; Suzuki & Rancer, 1994) and inconsistency in the understanding of metrics (e.g., Horn & McArdle, 1992), comparisons and conclusions drawn from such data are likely to be invalid and misleading. Data gleaned from LVC and distributed training environments have a high potential to produce violations of measurement invariance unless standards can be put in place to ensure consistency. As an example, imagine a P-8A simulator and a MH-60R simulator are performing a complete Anti-Submarine Warfare (ASW) mission in concert, where one MOP, *Sonobuoy Usage,* is the efficient use of sonobuoys. In this hypothetical example, the definition of *Sonobuoy Usage* for the P-8A considers the total number of sonobuoys dropped versus the number in contact with the target. The MH-60R in this illustration cannot publish the number of buoys in contact so they define *Sonobuoy Usage* as the total number of buoys dropped. In this scenario, making a valid comparison regarding which platform most efficiently employs sonobuoys is impossible with the existing metric. Any conclusions drawn are likely to be misleading. In order to completely and accurately understand whole force proficiency, it is imperative to define dimensions of performance and implement HPMs consistently.

Likewise, consistency of the feedback information provided to trainees in debriefing is crucial. While, many simulation-based training environments use playback technologies and instructor observations (Wiese et al, 2012), there are several weaknesses that could negatively affect learning and retention. First, training scenarios can last several hours, making it difficult for instructors to accurately remember specific instances of good and bad performance even when they are able to take time-stamped notes. In addition, this approach lacks analytical tools that ensure standardization and facilitate improved understanding of root causes of performance issues, particularly for inexperienced instructors. By facilitating timely delivery of feedback that is diagnostic in nature through HPMs, the feedback is more likely to result in retention producing a more effective and efficient training paradigm (Wiese et al, 2012).

Reuse of Standardized Measures

Another benefit of standardizing performance measures is the ability to reuse them with other platforms that perform the same or similar mission sets, saving development time and resources. As APMs become more commonplace in training environments, the development of a library of mission-based metrics would be a logical next step to support reuse across communities. An APM library should enable training communities to integrate already developed and standardized measures into their existing architectures, by-passing or reducing the need for costly development. As a practical example, the authors have explored the reuse of measures between the P-8A and MH-60R. For the P-8A community, the team defined measures for ASW missions, resulting in roughly 45 measures. Following this work, a small budget effort ensued to develop ASW measures for the MH-60R platform. To accelerate APM development, the team leveraged commonalities between platforms. The performance measurement concepts from P-8A where reviewed to identify where platform variations required measure rework to meet MH-60R needs. Current and future platforms interested in developing ASW measures for their training systems can increase speed-to-the-fleet with minimal resources by reuse, expansion, and/or tailoring of this ecosystem of APMs (Atkinson & Killilea, 2015).

Mechanism to Validate Observer-based Assessment with System-based Measures

While system-based measures provide us with information about what happened, observations can illuminate the behaviors and thought processes that led to these successful or unsuccessful outcomes. Currently, many training environments still rely on paper and pencil gradesheets to capture instructor observations and assessments. Incorporating both automated and observed measures of performance permits the construct validation of each. Validation of whether or not something is measuring what it purports to measure is accomplished by assessing its correlation with something that is conceptually related. As an example, a MOP easily calculated by a system may be *timeliness to reestablish contact with a target*—calculating the total time from when contact is lost to when it is regained. In order for aircrews to quickly reestablish contact, aircrews coordinate and communicate with one another, sub-facets of Crew Resource Management (CRM) that are observational measures of performance. If aircrews maintain high marks on observer-based measures of CRM, we would expect to see an associated high score with *timeliness to reestablish contact*. If there is not a high correlation, it is evidence that one or the other measure is either an unreliable and/or invalid measure.

Understanding Training Trends

At the event level, APMs provide instructors with valuable standardized data for debriefs and for preparation for upcoming training events. For decision makers analyzing the outcomes of multiple events, the results illuminate trends that may have implications for modifying training curriculum. First, identifying skills trained in one environment that are not transferring to another indicate training system and/or curriculum issues. Second, results of APMs can support streamlining the training cycle by providing a performance driven, as opposed to qualification driven, training readiness curriculum. As an illustration, qualifications training syllabi are based on completing a specific number of training iterations for each mission a platform performs. Use of APM data would allow trainees and aircrews who have achieved an acceptable level of proficiency on one mission set or skill set (e.g., tactical skills) to progress to the next, rather than perform redundant training. Finally, a data science approach to performance trends may yield predictive analyses that could highlight how to utilize resources in different mission profiles.

DEVELOPING A STANDARD FOR HUMAN PERFORMANCE MEASUREMENT

A SISO Product Development Group (PDG) is developing the first version of the Human Performance Markup Language (HPML) standard; a full description of the HPML schema can be found in the HPML user guide posted on SISO. HPML relies on Extensible Markup Language (XML) to provide schemas for organizing the information relevant to defining performance measurements, including computations, measures, assessments, results, and instances/periods (Walker, Tolland & Stacy, 2015). Since the inception of HPML (Stacy, Merket, Freeman, Wiese, & Jackson, 2005; Stacy, Merket, Puglisi, & Haimson, 2006), the goal has been to meet the needs of a variety of services and domains (Walker et al., 2015) by providing an intrinsic generalizability. At a high level, HPML provides a means for configuring and executing measurement and assessments of performance through a flexible framework (Atkinson & Killilea, 2015). The HPML standard is an XML/JSON-based language designed to express performance measurement concepts that are machine and human readable to increase ease of use and accessibility.

At the basic level, performance measurement instructions defined in HPML specify the system data elements to be collected, the calculations to use to process the data, and when to produce performance measurement results. At a high level, HPML is composed of different sub-schemas that each represents different parts of the standard. These modular components contain dependencies that work together to calculate MOPs. Figure 1 provides a schema dependencies diagram that shows the links between each of the main groups within the schema. The schema is separated into six distinct groups, 1) Data, 2) Computation, 3) Results, 4) Assessments, 5) Measures, and 6) Instances and Periods. These groups make up the core components of HPML and can be added to or expanded with additional links in the schemas.

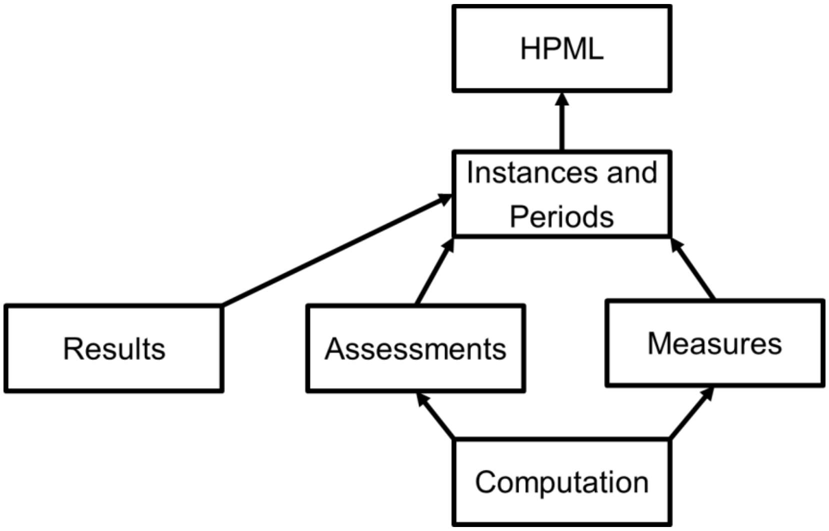

Figure 1. Schema Dependencies Diagram

Data

Data refers to the schemas that represent the linkage between data sources (e.g., HLA, DIS) and computation. This schema provides an abstraction to select data points from various data streams and sources. For instance, you could select a world location as a relevant aspect for a measure (e.g., proper weapon placement). In NASMP this data point may be found in a different location than it is in DIS. HMPL *Data* provides the flexibility to switch between NASMP and DIS. With minimal rework, *Data* defines the raw data sources for use in the *Computation* schema.

Computation

Computations refer to the schemas that represent the algorithms, triggers, and other computational components of HPML that provide input for *Assessments* and *Measures*. *Computations* provide a simple way to define how a measurement or assessment should be calculated by defining which data points should be combined in an algorithm to produce a measurement. Users can expand *Computations* through scripts which can be defined in the user's language of preference (e.g., lua, cscript, R, python) to allow executable code to know how to handle calculations.

Results

Results refer to schemas representing the output of both measures and assessments, detailing the information produced by the contributions of *Data* and *Computations*. *Results* are an important aspect of human performance because it allows you to see the entire context of the result in a human-readable way. This may include information such as who was involved, what assessments were assigned, and for complex measures the sub-facets of performance that were fed into the data. The resulting structure can be traced backwards to see all results in the tree that contributed to the final top-level result, providing you with a comprehensive history and contextual information.

Assessments

Assessments refer to schemas representing the labels given to measures either by category (e.g., Expert, Novice) or value (e.g., 99%, 75%). While not all APMs require *Assessments*, this aspect of APM provides information that allows instructors to quickly understand the *Results* output of measures.

Measures

Measures refer to schemas representing the linking of data sources and computations to produce measurement outputs from a given data source. Measurement provides the method to connect those *Data* from specific sources and *Computations* to build a result and possibly an *Assessment*. *Measures* also provide a way to link multiple simple measurements together to provide complex or higher level measurements of teams and teams-of-teams.

Instances and Periods

Instances and Periods refer to schemas representing the creation and use of measures and assessments for a given context. This schema defines the instantiation of HPML elements at specific points in time (e.g., phase of a mission), or specific locations within space (e.g., variation in performance for different areas of operation). Every element in this schema has a time and/or location component.

HPML Status

The HPML standard is moving towards a version 1.0 release of the HPML standard through the SISO PDG, which is made up of researchers and engineers from different fields of expertise. The PDG uses SISO discussion boards and working groups to collaborate and decide on ways to improve the standard. Currently, the HPML PDG meets twice a year at SISO conferences to discuss and expand upon the standard.

Some examples of activities initiated by the SISO PDG include developing means to increase interoperability with other performance schemas like xAPI, providing more ways to define computation via scripts, and defining competency models. As part of the HPML PDG activities, xAPI was reviewed for similarities and differences. Ultimately the PDG decided that the main difference between the two was HPML’s ability to define how the measurement and assessment is performed. This is primarily related to the computation, data, measurement, and assessment subschemas. While xAPI provides a way to present performance, it lacks a mechanism to identify performance from raw data. In order to facilitate better interoperability, a set of guidelines was developed to translate HPML results into xAPI statements. This provides the end user with more ways to incorporate HPML into their existing systems. Additionally, the group is working to expand the computation schema to include other popular scripting and algorithm languages to support the incorporation of existing models and algorithms into HPML to expand the functionality. Finally, in order to further the understanding of the team or individual performance, the PDG has identified the need to develop an extendable competency model in HPML. This new subschema will enable the end user to link existing models with HPML measures and assessments to gain a better understanding of the human.

CONCLUSIONS AND FUTURE DIRECTIONS

As previously noted, leveraging big data sources for measuring human performance in training has the potential to provide assurance that systems provide effective training (Wiese et al, 2012). Moreover, as large scale integrated training environments such as LVC increase, APMs are an analytic tool that will enable feedback and trend analysis. As a result of data driven decision making, APMs can provide a means to ensure a higher standard of training and ultimately a better warfighter.

While the value of using data analytics to aid in HPM is still in its infancy, ensuring current and future training development efforts make appropriate considerations for implementing APM remains a challenge. The development of an industry/government standard would provide guidance to developers about the data that should be published in order to generate assessments conducive to instruction. Aside from this, a standard along with associated measure libraries have the potential to help reduce development, sustainment, and maintenance costs as APMs are integrated into training systems. Further, the expectation of distributed training is that aircrews will walk away with the knowledge, skills, and tactics necessary to be highly effective, integrated warfighters. The only way to know with any certainty that we are achieving this aim is to measure performance.

However, the effectiveness of measurement requires reliable and valid MOPs. Specifically, MOPs and resulting feedback must be accurate and consistent from one training session to the next if they are to result in higher performing aircrews. One of the benefits of APMs, in contrast to observer-based measures, is their consistency between instructors. Their definitions and metrics do not run the risk of being misinterpreted or biased in any way, resulting in exceedingly high reliability and improved validity. As noted within this paper, inconsistencies in data published currently result in systems contributing various amounts of data relevant to understanding performance, and unless requirements driven by standards are established system developers will maintain the status quo resulting in limited success for HPM.

ACKNOWLEDGEMENTS

The views expressed herein are those of the authors and do not necessarily reflect the official position of the DoD or its components. Sponsors for underlying research and development that have informed the HPML proposed standard and use cases from a Navy perspective have included the Naval Air Systems Command (NAVAIR) PMA-205 Air Warfare Training Development (AWTD) program, PMA-290, the Small Business Innovative Research/Small Business Technology Transfer (SBIR/STTR) program, Office of Naval Research Rapid Innovation Fund (RIF), and the Naval Innovative Science and Engineering (NISE) program (Section 219).

REFERENCES

Astwood, R. S., Van Buskirk, W. L., Cornejo, J. M., & Dalton, J. (2008). The impact of different feedback types on decision-making in simulation based training environments. *Proceedings of the Human Factors and Ergonomics Society 52nd Annual Meeting, 52*(26), 2062-2066.

Atkinson, B. F. W., & Killilea, J. (2015). A review of the potential return on investment benefits of a human performance measurement standard: Lessons learned in the navy aviation community. *Proceedings of the Fall Simulation and Interoperability Standards Organization (SISO) Interoperability Workshop*, Orlando, FL.

Blair, D. (2015). Moneyjet. *United States Naval Institute Proceedings. 141*, 68-70.

Buss, D. H. (2014, April). Naval aviation vision: A preeminent warfighting force, today and in the future. *Aviation, Inside the Navy*. Retrieved from http://navylive.dodlive.mil/2014/04/16/naval-aviation-vision-a-preeminent-warfighting-force-today-and-in-the-future/.

Buss, A. R. & Joyce, J. R. (1975). Ontogenetic changes in cognitive structure from a multivariate perspective. *Developmental Psychology, 11*, 87-101.

Horn, J. L. & McArdle, J. J. (1992). A practical and theoretical guide to measurement invariance in aging research. *Experimental aging research, 18*(3), 117-144.

Irvine, S. H. (1969). Factor analysis of African abilities and attainments: Constructs across cultures. *Psychological Bulletin, 71*(1), 20-32.

Kluger, A. N. & DeNisi, A. (1996). The effects of feedback interventions on performance: A historical review, a meta-analysis, and a preliminary feedback intervention theory. *Psychological Bulletin, 119*(2), 254-284.

MacMillan, J., Entin, E. B., Morley, R., & Bennett Jr, W. (2013). Measuring team performance in complex and dynamic military environments: The SPOTLITE method. *Military Psychology, 25*(3), 266-279.

McAfee, A. & Brynjolfsson, E. (2012). Big data: The management revolution. *Harvard Business Review, 90*(10), 60-68.

Naval Aviation Enterprise. (2014). *Naval aviation vision: 2014-2025*. NAE Publication Distribution.

Naval Aviation Enterprise. (2016). *Naval aviation vision: 2016-2025*. NAE Publication Distribution.

Portrey, A. M., Keck, L. B., & Schreiber, B. T. (2006). *Challenges in developing a performance measurement system for the global virtual environment* (Report No. AFRL-HE-AZ-TR-2006-0022). Air Force Research Laboratory, Warfighter Readiness Research Division, Mesa, AZ: Department of Defense.

Schleicher, D. J., Day, D.V., Mayes, B. T., & Riggio, R. E. (2002). A new frame for frame-of-reference training: Enhancing the construct validity of assessment centers. *Journal of Applied Psychology, 84*(4), 735-746.

Simulation Interoperability Standards Organization (SISO) Distributed Debrief Control Protocol (DDCP) Study Group. (2011, April). *Final report for: The distributed debrief control protocol study group* (technical report SISO-REF-028-2011). SISO.

Stacy, W., Ayers, J., Freeman, J., & Haimson, C. (2006). Representing human performance with human performance measurement language. *Proceedings of the Fall Simulation Interoperability Workshop*, Orlando, FL.

Stacy, W., Freeman, J., Lackey, S., & Merket, D. (2004). Enhancing simulation-based training with performance measurement objects. *Proceedings of the Interservice/Industry Training, Simulation, & Education Conference*, Orlando, FL.

Stacy, E. W., Merket, D., Freeman, J., Wiese, E., & Jackson, C. (2005, December). A language for rapidly creating performance measures in simulators. In *Proceedings of the 20005 Interservice/Industry Training, Simulation & Education Conference* (pp. 207-218).

Stacy, W., Merket, D. C., Puglisi, M., & Haimson, C. (2006). Representing context in simulator-based human performance measurement. *Proceedings of the Interservice/Industry Training, Simulation, & Education Conference*, Orlando, FL.

Suzuki, S. & Rancer, A. S. (1994). Argumentativeness and verbal aggressiveness: Testing for conceptual and measurement equivalence across cultures. *Communications Monographs, 61*(3), 256-279.

Walker, A., Tolland, M., & Stacy, W. (2015). Using a human performance markup language for Simultor-based training. *Proceedings of the Fall Simulation and Interoperability Standards Organization (SISO) Interoperability Workshop*, Orlando, FL.

Wiese, E., Atkinson, B. F., Roberts, M., Ayers, J., & Ramoutar, D. M. (2012). Automated human performance measurement: Data availability and standards. *Proceedings of the 34th Interservice/Industry Training Simulation & Education Conference*, Orlando, FL.

Wiese, E. E., Nungesser, R., Marceau, R., Puglisi, M., & Frost, B. (2007). Assessing trainee performance in field and simulation-based training: Development and pilot study results. *Proceedings of the Interservice/Industry Training, Simulation & Education Conference* , Orlando, FL.