Standardizing Performance Measurement While Ensuring Psychometric Validity

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

ABSTRACT

Previous research and development work by the Navy has focused on developing an industry standard for system generated performance measurement that facilitates the mining of individual operator and aircrew performance data from simulators as an effort to continue the advancement of training systems. However, a standard alone does not provide the guidance necessary for implementation that will ensure this measurement medium adheres to psychometric principles. A major focus of science is ensuring reliability and validity when measuring psychological constructs. That is, when we measure a psychological construct (e.g., intelligence) is there consistency (i.e., reliability) within and between measures (e.g., observers, computer systems), and are we measuring what we intend to measure (i.e., validity). While this latter distinction may seem obvious, people often make the mistake of measuring something that is unintended (i.e. construct contamination) or not fully measuring what they intend to measure (i.e., construct deficiency), and then use that information to inform feedback and decisions. The consequences of these measurement mistakes may be critical. As we work toward a standard that guides engineers in the development of system generated measures of performance, this standard must also incorporate important psychometric concepts and analytical techniques such as reliability, validity, local data norming, regression weighting, and structural equation modeling to name a few. One of the more robust findings in psychology is that ratings made by humans inherently contain some degree error. This error reduces the usefulness and legitimacy of findings. While system generated measures of performance will be void of these common human errors, they will still need to adhere to psychometric principles to ensure their utility in applied and academic settings. This paper will provide a review of the science of human performance measurement and provide specific recommendations for policies and approaches that support the automated performance standard.

ABOUT THE AUTHORS

Mitch Tindall, PhD, is a Research Psychologist at Naval Air Warfare Center Training Systems Division in the BATTLE Laboratory. During his graduate work, he performed research and consulting in human performance and productivity enhancement. His work for the Navy includes several areas such as psychometric development and validation, human performance measurement, HCI, data management and analytics, training systems enhancement and validation, and systems software evaluation. His research interests include human performance measurement, physiological event training enhancement and big data analytics. His Ph.D. is in Industrial-Organization (I-O) Psychology from University of Central Florida (UCF).

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.

Standardizing Performance Measurement While Ensuring Psychometric Validity

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

As a means to understand how one can enhance performance in professional and/or occupational settings, the discipline of psychometrics seeks to improve the validity, or accuracy, of ratings (e.g., Alcazar, Hopkins, & Suarez, 1986; Bobko & Colella, 1994; Pritchard, 1990). This focus on improving validity is because, traditionally, performance is rated by a knowledgeable observer (e.g., manager, supervisor, coworker, commanding officer). Unfortunately, these human assessors rate others' behavior with a degree of error (Viswesvaran, Schmidt, & Ones, 2005). This is problematic, as the amounts of errors in a rating of performance increases, the ability to be confident of any result of the research or to derive value of the feedback from the flawed rating decreases. Research efforts intended to reduce error and improve the accuracy of performance ratings have included, rater error awareness training (Bernardin & Pence, 1980; Woehr & Huffcutt, 1994), frame-of-reference training (Guion, 1998; Hauenstein, 1998), 360 degree ratings (Dalessio, 1998), inclusivity in assessment formation and design (Latham, Mitchell & Dossett, 1978), to name a few. Common to all these interventions is the motivation to improve the validity of human rated performance. More recently, technology maturation has increased performance assessment opportunities through development of system-based, automated performance measures and integration of system-observer performance databases to track longitudinal trends. While these technologies provide rich data sets, they are also subject to validation challenges that must be considered. That is, without standards and policy in place to guide implementation, retaining valid measures may be difficult to manage.

PSYCHOMETRIC VALIDITY IN THE AGE OF SYSTEM-GENERATED PERFORMANCE MEASUREMENT

"It is widely believed, along with significant support through research and case studies, that organizations that use data to make decisions over time in fact do make better decisions, which leads to a stronger, more viable business." (Dean, 2018, p. 20). Fortunately, human observation and assessment is no longer the only method for rating performance. The proliferation of simulation-based training occurring in military contexts creates an incredible opportunity to mine systems for data relevant to human performance. In these contexts, where simulation handles an ever-growing proportion of fleet training, data from systems can be mined to attain ratings of select dimensions of performance and do so without the error humans are prone to making. Additionally, these system-generated measures of performance (MOPs) can assess *processes* (i.e., behaviors and cognitions) that led to an *outcome* (i.e., success or failure) in addition to the outcome itself. This is an important point of emphasis because past attempts to mine systems often resulted in outcome data only (i.e., revenue, profit, number of accidents, timeliness) and little if any process data (i.e., communication, coordination, organization, use of knowledge, problem solving, decision making). Process data is incredibly valuable as it illuminates the cognitive and behavioral processes that led to a subpar, par, or exceptional outcome. Process data/information is imperative for effective training interventions as it provides important detail useful for identifying gaps and inspiring innovation. Furthermore, when process information is used as feedback it is more likely to improve subsequent performance than outcome-oriented feedback (Earley, Northcraft, Lee & Lituchy, 1990). While system-generated data shows promise for advancing the science and applied use of human performance measurement, it is important to understand its limitations and how some of those can be addressed.

Past theory and science can be applied to help ensure the validity of system-generated MOPs by advocating for specific standards and policies that could be adopted by organizations within industry and/or the Department of Defense (DoD). Specifically, standards provide a formally approved product that reflects consensus across organizations in the form of agreements to products, practices, or operations to ensure interoperability. For this domain, Human

Performance Markup Language (HPML), currently under consideration as a Product Development Group within the Simulation Interoperability Standards Organization (SISO) is one such standard. However, at this time there are no known policies in place or under development that would ensure the use of such standards and guide mechanisms to support implementation. This paper outlines an existing literature base that can be used to provide guidance for policy development and outline an architecture that integrates a scientific approach within a broader semi-automated performance measurement capability.

It is important to note that the theoretical and analytical techniques discussed in this paper are not novel and have been used extensively as research methods in psychology. These methods have informed experimental and quasi-experimental designs as well as advanced theory. In select cases, they have been used in applied settings (Pritchard, Harrell, DiazGranados & Guzman, 2008) to improve objectivity and apply scientific methodology to people in work settings. As the proliferation of systems capable of assisting us in assessing performance proceeds, it is crucial that we consult past theory and empirically validated research in the area of human performance measurement. This system-centric approach to assessment should now evolve alongside the science, one informing the other. A motivation of this paper is to raise awareness in academia about the current state of system-based performance measurement and to excite the field about what is possible. Additionally, while this paper conceptualizes how traditional methods used in the science should inform standards, policies and architectures, it is by no means an exhaustive list. The paper should provoke intellectual discourse between academics, applied practitioners, and software engineers regarding how other novel methods could inform policy and standards and assist the continual advancement of system-based HPM.

THE IMPORTANCE OF TAKING A CRITERION CENTRIC APPROACH (CCA) TO VALIDATION

Criterion as defined in the field of Industrial Organizational Psychology is synonymous with the dependent variable or variable we are interested in predicting. For the purposes of this paper, the criterion is performance, which is often the main predictor for businesses, government institutions, and academia. Naturally, we want to know what tools or information will help us predict and improve performance. With such a heavy emphasis on a single variable, it is crucial that this variable is measured accurately, and that all stakeholders agree to the aspects of performance that are most important, as they will impact selection, placement, and training personnel.

Bartram (2005) identified a process (see Figure 1) by which a criterion is defined, agreed upon, validated before other human resource processes (e.g., recruitment, selection, training, compensation) are identified and implemented. The idea is to empirically define what you are predicting and develop your metric before you choose your predictors and methods for training or enhancing performance. As we continue to advance system-generated MOPs, it is important we consider Bartram's CCA as a standard for developing MOPs. Unlike other analytical techniques and methods discussed throughout this paper, CCA presents steps and factors relevant for human performance assessment policies. Specifically, the steps within CCA are important so MOPs are not selected arbitrarily and to narrow your list of MOPs to what matters most.

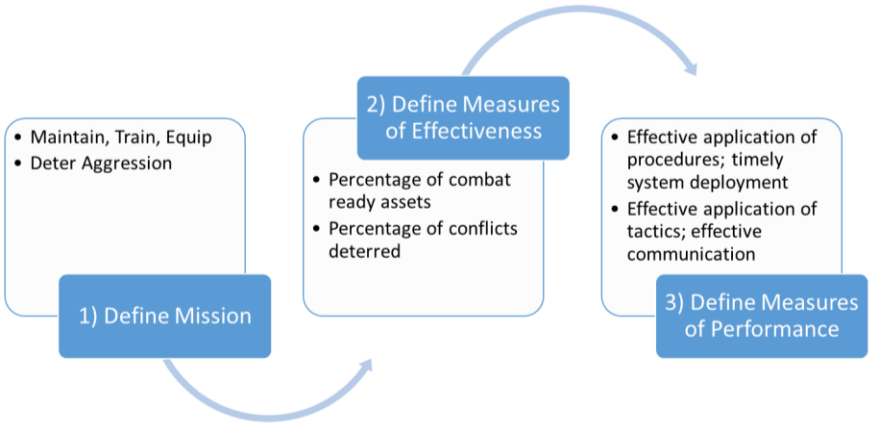

Figure 1. Example of CCA to Performance Measure Development Derived from U.S. Navy Mission

Step One: Defining the Mission

The first step in developing criterion according to Bartram (2005) is to define the overarching goal(s) of the organization and to have leadership agree on that goal(s). This can be derived from an organization's mission statement or be the consensus of leadership. The point of defining the overarching goal is to ensure alignment between the mission and subordinate outcomes and processes. That is, all measureable outcomes and the individual and team processes should, to some extent, affect the overall mission. This ensures the organization is not measuring or doing things that are not productive or conducive to meeting its goal(s). It is important to note that missions can be defined at lower and higher levels of organizations such as the DoD, specific services, or platform squadrons. To demonstrate the practical application of CCA for system-generated MOPs we will use the US Navy as an illustration (Figure 1). The mission of the U.S. Navy (USN) is "to maintain, train and equip combat-ready Naval forces capable of winning wars, deterring aggression and maintaining freedom of the seas." (Navy.com, 2018). To determine the degree with which we are meeting this mission, outcomes or measures of effectiveness are needed.

Step Two: Defining Measures of Effectiveness

The second step in developing criterion is to identify measures of effectiveness (MOEs) or outcomes that tell you whether you are meeting your mission based on outcomes that are easily measureable. Based on critical aspects of the USN mission, examples would include combat ready assets, the percentage of proficient fleet personnel, and the percentage of aggressive acts successfully deterred. Identifying and defining MOEs support determination of whether or not you are meeting your mission.

Step Three: Defining Measures of Performance

The third and final step in the CCA is to identify and define your MOPs derived from MOEs. MOPs are the individual or team behaviors and cognitive processes that lead to an outcome (Bartram, 2005). Using the same example, these could include communication, coordination, problem solving, knowledge application, and understanding of procedures.

Mitigating Cost and Maintaining Speed-to-Fleet While Being Criterion Centric

While standards are under development to ensure that robust performance measures are implemented in a manner that is consistent, reusable, and interoperable (Atkinson, Tindall, Tolland, & Dean, 2017), policies to guide implementation of these standards and resulting technologies should incorporate principles of CCA. For example, this means involving stakeholders from the top and bottom of the chain-of-command in the development process. Additionally, work on MOPs should not begin before our mission is defined broadly and MOEs are identified and defined. This will ensure a more efficient focused approach to measure development where all MOPs and MOEs affect the organizations broader mission.

The current state-of-practice for measure development, makes facilitating a true CCA practically difficult. After the software system architecture is developed by engineers, specific measures must be hard-coded into the architecture. This means every change to an existing measure, implementation of a new measure or elimination of an obsolete measure requires software engineers to program those changes. Further, if resources do not exist for continued maintenance of measures, they will become outdated or unused. While technology is under development to help ease the burden of measurement sustainment, proper implementation guidance is another policy consideration.

APPLYING PSYCHOMETRIC PRINCIPLES TO THE DEVELOPMENT AND REFINEMENT OF SYSTEM-GENERATED MOPS

Human cognition and behavior are incredibly complex making measurement an exceptional challenge. Unlike much of the data in physical sciences, this psychological phenomenon is always measured with a degree of error. That is, modes of measuring psychological constructs are either *contaminated*, measure something other than the construct of interest, and/or *deficient*, the inability of a mode of measurement to measure the whole construct (MacKenzie,

2003) (See Figure 2). As a rule, the greater the proportion of error in your data the less likely you will be able to find an effect. Additionally, inferences made from this error prone data are likely to be flawed. As a result, a principal focus of psychometrics is decreasing the degree with which psychological constructs (e.g., intelligence, personality, performance, emotions) are measured with error. Two ways a Psychometrician can improve the measurement of psychological constructs is by analyzing and improving the reliability and validity with which constructs are measured. This process is known as construct validation.

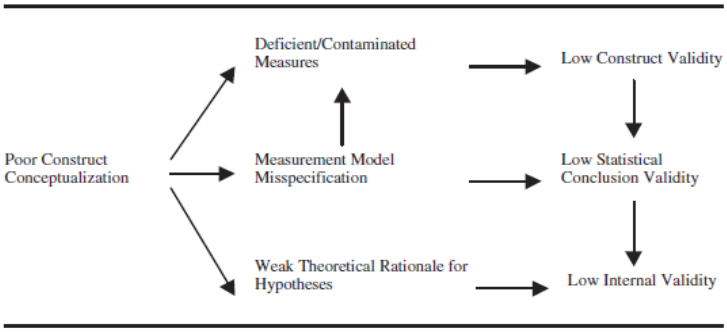

Figure 2. "Consequences of Poor Construct Conceptualization" (MacKenzie, 2003, p. 324)

Reliability is the consistency of scores obtained from a single individual when they are assessed multiple times or with collections of equivalent items (Anatasi & Urbina, 1997), and sets the upper bound for validity. This means a single measure can only be as valid as it is reliable. In other words, if a measure is reliable, individuals with the same level of knowledge, skill, or ability should produce similar results. Reliability is assessed one of several ways (e.g., test-retest, inter-item/Kuder-Richardson reliability, parallel forms or split-half reliability) and is analyzed using correlation coefficients such as the Pearson Product-Moment (Pearson, 1907), Spearman-Brown (rho) (Kelley, 1925) and Cronbach Alpha (Cronbach, 1951). Simply put, higher correlations (> .80) are an indicator of acceptable reliability. Factors such as the size, type, and dimensionality of your data should be considered when determining which approach to assessing reliability makes the most sense for a specific use case.

Validity pertains to the degree to which you measure what is intended (Anatasi & Urbina, 1997). As an illustration, if I develop a test to measure '*math skill*', I want to be confident that the test is indeed, measuring '*math skills*' and not '*comprehension*'. Validity is less straightforward than reliability. It is assessed in several ways (face, content, construct, and criterion) but in general examining correlations of the measure with theoretically related (i.e., convergent validity) and unrelated (i.e., divergent validity) phenomena will provide evidence for validity. For example, for our math test to be considered valid the test should be related to math grades in school, occupational choices in the math fields, and scores on other mathematics tests. It is important to note that an item can have high reliability but still lack validity. This occurs when a measure we develop is consistently measuring something that is not intended. For example, personality assessments are reliable and valid measures of personality traits. However, when they are administered as part of a job selection process they tend to measure people's perceptions of employers expectations. That is, in the context of work, people do not fill out personality measures based on their personalities but instead on the behavior they think the organization expects of them. This point is important as the proliferation of system-based measures of performance continues. System-based measures of performance will, inherently, maintain high reliability, but there will still be a concern about validity that demands attention.

As illustrated in Figure 2, the consequences of invalid measurement are serious. In military contexts, those consequences have life and death implications. Leadership now firmly acknowledges the value of making objectively informed decisions. To be truly objective, one must forgo intuition and instead rely on facts to inform decisions. However, this only works if those facts are accurate. This is the essence of validity. As soon as a mistake occurs because of relying on misleading or false facts, we lose confidence in the system and regress to making decisions based solely on intuition and emotion. As shown in prior research, this increases the likelihood of an incorrect decision being made. (Dean, 2014). It is for this reason we advocate for policies that require, at minimum, an abridged approach to measure validation for system-based measures. Specifically, a policy for deleting obsolete measures, modifying existing measures and developing new measures should be implemented to ensure the validity of our data. The policy or instruction should require the right mix of domain experts, leadership, and scientists to facilitate the validation process and sign-off on final measures before implementation. Doing this with large data sets, like those common in military settings, could be a huge undertaking. Fortunately, advances in technology can be leveraged to make the process more manageable.

Leveraging Technology to Automate the Psychometric Validation Process

Full-scale validation studies are time consuming and costly endeavors; however, automation, machine learning, artificial intelligence and centralized data storage affords us the ability to build validation into our performance measurement systems. While tools that enable users to update or build new MOPs saving resources and increasing currency with evolving tactics, there is a risk that measures will be developed without the scrutiny necessary to ensure reliability and validity. Implementing an architecture that leverages a central storage of performance data provides an opportunity to use algorithms to automate the process of testing measures reliability and validity. In addition, those algorithms could periodically or continuously assess the validity of existing measures alerting users that modifications need to be made to maintain its validity.

For reliability, equations (e.g., Pearson Product-Moment, Cronbach Alpha and Spearman-Brown) can be translated into algorithms that automatically assess the reliability of MOPs. Once reliability has been established, the system could be used to automate the assessment of validity. If instructor grade sheets and post-mission data are stored on the same servers as the system-generated MOPs, that data could be retrieved to assess construct and criterion validity. In this case, algorithms that apply coefficient alpha could be used to assess the correlation between conceptually similar phenomena. For construct validity, one could argue that instructors' grades of an aircrews 'Crew Resource Management' should be related to system-generated 'Timeliness of Attack' because crews that coordinate more effectively should be able to engage more quickly. If the resulting Cronbach Alpha between these two measures is high (>.80) it is an indication that one or both are valid. Criterion validity can be assessed by retrieving post-mission data and relating it to your system-generated MOPs. This is also an excellent way to assess the entirety of your measurement system. For example, a system-generated MOP that produces an overall score should be related to certain post-mission data such as the time it takes to gain contact on a target, the number of expendables used during a mission, accuracy of target classification and the like.

Automating the psychometric validation of performance measurement systems is not only useful for ensuring the validity of the measurement system it can also help leadership gain insight on the return on investment (ROI) for new training interventions, changes to TTPs, advances in technology, and any intervention expected to have an impact on performance of an aircrew. Automating performance measurement is an initial step in improving human performance assessment as a whole. However, if we are to automate measurement we should also take the next step and automate the assessment of measure's reliability and validity to ensure system utility and quality and save time and resources in the long run.

UPDATING ASSESSMENTS BY AUTOMATING THE LOCAL NORMING OF DATA

An assessment in this context, is generally defined as a schema representing labels given to measures either by category (e.g., expert, novice) or value (e.g., 98%, 50%) (Atkinson, Tindall, Killilea, Tolland & Dean, 2017). Put another way, assessments provide meaning (e.g., labels, percentages, percentiles) to raw data. Current best practice for assessing raw data requires subject matter experts (SMEs) to decide, arbitrarily, what constitutes mastery or poor, average, and good performance. While this approach is easy to implement, it often reflects the biases held by your SMEs and thusly be inexact. For example, a severity bias can occur when SMEs have made assessments too high and unreachable by trainees (Fischer & Jungerman, 1996; Franic & Pathak, 2000; Weber & Hilton, 1990). Such a bias can have the unintended effect of demotivating trainees. On the other hand, SMEs may also possess a leniency bias that results in assessments that are too easy (Borman, 1975). These are only two, of many, biases that can impact the accuracy and utility of assessments when they are solely derived from SMEs.

Timeliness of Attack	
Percentile Ranking	Raw Scores (minutes)
10	90
20	85
30	80
40	75
50	70
60	65
70	60
80	55
90	50

Figure 3. Notional Example of Local Norms

Given you have enough data, the best way to form assessments to ensure accuracy and utility is to use local norms (Elliot & Bretzing, 1980). This is a data driven approach that allows the performers themselves to set the bar. A classic example of the use of local norms is in the classroom. Professors will build difficult tests and use a curve where the highest

performing student sets the bar. In this example, the highest performing student may have received a raw score of 80/100 on the exam. Adding a curve means the test is now out of 80 and not 100.

Local norms work much the same way. Once you have collected enough data (N > 200) from your population of interest you simply attain the mean or your 50th percentile. Each standard deviation away from the mean would add or subtract 10%. Developing assessments this way is incredibly straightforward your average performers will be at or within one standard deviation of the mean, your novices would be at least two standard deviations below the mean and experts at least two standard deviations above the mean (See Figure 3).

The use of local norms has several theoretical and empirically supported benefits:

1. Local norms narrow your focus on the population of interest and help you deal with range restriction.
2. The use of local norms eliminates biases associated with arbitrarily formed assessments reducing the possibility of general and legal disputes.
3. Local norms are inherently fair as it is the trainees at the same levels of experience that set the assessments not experts with ample experience.

The Special Importance of Local Norms for System-Generated Data

One could argue that the use of local norming to formulate assessments is more relevant in environments that use system-generated data like military domains. These domains are able to constantly collect and analyze data that can easily be fed back into norms. This is important because military domains are highly dynamic. What constitutes expert performance yesterday may not be what constitutes expert performance today. Because the collection of this data is already happening, we advocate for policies that rely on that data to set assessment standards based on local norming of the data. This not only ensures the accuracy and utility of the data it also has the added benefit of not requiring time by SMEs to determine values and categories for assessments. Tools that automate local norming of data, again, can be built into the broader performance measurement system to automate the process and provide continuous updating.

CONCLUSIONS AND FUTURE DIRECTIONS

None of the analytical techniques traditionally or currently utilized in the field of human performance measurement as outlined in this paper are novel, but the application of them with this new medium (system-generated data) is innovative. The notion of Moore's Law implies that computational processing speed doubles every eighteen months or simply that technological advancement in the computer age happens at an exponential rate (Schaller, 1997). If this maintains true for computers and technology that facilitate system-generated performance data then we may, yet again, allow the capability to outpace the science resulting in dire consequences for the warfighter and the military as a whole. Performance measurement systems and resulting data should be developed and maintained with science in mind to ensure the accuracy and utility of the resulting system-generated performance measurement structure. Doing this at the early stages of the technology development will save immense amounts of time and resources in redesign and redevelopment costs when the systems we generate do not work as they are intended. These costs do not include the human, hardware and infrastructure costs when accidents inevitably occur as warfighters engage in behaviors that result from misleading or outdated performance measures. It is important to note that performance measures used in training not only help us assess proficiency, they help us encourage desired behavior and discourage undesirable behavior.

The psychological sciences are constantly evolving and new techniques are brought to the forefront to facilitate the continual advancement of the field as a whole. Therefore, this paper should be seen as a starting point for an ongoing discourse between engineers, statisticians, and scientists about how system-generated performance data can help advance the science and how the science can assist in ensuring the validity and quality of system-generated performance measures in applied settings. For example, one area that is seeing advancement is how to deal with large data sets in environments that change rapidly (e.g., military context).

Applying Novel Analytical Techniques to Deal with Highly Dynamic Data

From an analytical standpoint, it is important to note the differences between human research based on experimental designs and those derived from big data streaming environments. Experimental designs have a finite beginning and

end. Additionally, researchers have more control over the independent (i.e., manipulated variable) and dependent variables (i.e., measured variable) and the context within the data collection environment. For example, if I wanted to understand the effect of a new training intervention (IV) on performance (DV), experimental designs provide confidence that the training intervention impacted performance and not some other variable (e.g., characteristics of study participants, contextual factors). Essentially, experimental designs enable the research isolate the specific effect of an IV on a DV.

Unfortunately, human research derived from big data streaming environments lack the control of experimental designs. They do not have a finite beginning or end. The context in which the data is collected is dynamic making it possible to either remove obsolete variables or add new variables that may impact your DV. Using the example above, if I am to rely on system-generated data to determine the effect of a training intervention on performance it may be difficult if not impossible to isolate the effect of the training intervention on performance outside of other variables (e.g., changes to TTPs, advancements to technology, characteristics of subjects). An analyst with extensive experience and education in statistics would be needed to mine and parse the data to determine, with confidence, the specific effect of the training intervention. The time and resources spent by such an analyst would be immense and this does not account for manually updating the database regularly whenever variables and the environment change. Recent advances in big data analytical techniques, artificial intelligence and machine learning algorithms, present an opportunity for dealing with the dynamic nature of big data in military contexts. Some of these techniques were developed to deal with the vast data synthesis required for cyber security while others for the challenges associated with the effective integration of data in the field of medicine (Dean, 2014).

Considering the aforementioned advances, the US Navy has recently begun a research and development (R&D) effort leveraging the Small Business Innovative Research (SBIR) program to deal with the issue of analyzing human data in dynamic contexts. Techniques to Adjust Computational Trends Involving Changing Data (TACTIC-D) seeks to develop a program that can detect change in data, automate the updating of existing models and provide easy-to-interpret visualizations to ensure stakeholders have situational awareness regarding those changes to the environment. The motivation of the effort comes from an understanding that big data sets will likely be used to assess trends overtime. This trend analysis will be useful for determining the ROI for past efforts, providing data to inform future innovations and interventions, and diagnosing problems in real-time as they arise. However, without technology or an analyst constantly detecting change, adjusting models and providing easy-to-interpret visualizations pertaining to analyses, it is unlikely data will be used or used effectively.

We are at the early stages of R&D efforts dealing with the changing nature of big data. As a result, we are not advocating for specific policies or standards in this paper. The outcome of this SBIR is expected to help inform future policies and standards regarding how we deal with changing data. Existing and future efforts intended to mine system-data for human performance information need to be aware of this issue and again technology should be leveraged to help automate and manage the process.

Figure 4 is a proposed framework for a performance measurement system that automates analytical techniques that ensure the validity, consistency, accuracy, and detail of measures of individual and crew performance. In addition, the architecture supports the continual updating of underlying models pertaining to performance and improves situational awareness of stakeholders interested in analyzing trends over time and across Squadrons, Wings, Platforms, and time. Stakeholders using this system to inform decisions will have a clear understanding of all the contributing factors to performance. This should lend greater confidence in decisions informed by this data as well as providing more detail to inform effective innovative ideas. Moreover, the automation of normative assessments will provide consistency between instructors and allow them to encourage crews to push the bar higher.

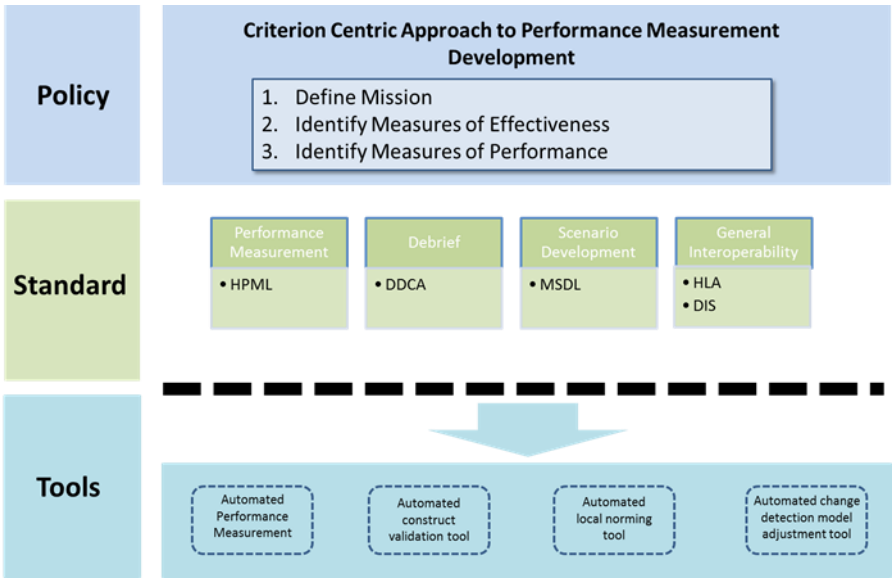

Figure 4. Notional framework for the integration of policies, standards and automated tools for the enhancement of performance measurement systems.

Ultimately, this paper stresses the need for policies that provide guidance on standard implementation and technology development to ensure the large amounts of data we are mining and analyzing is done in a scientific manner. Such an approach ensures the effective use of data. While the tools embedded in the above framework are far more complex than insinuated by the model, they represent a first step toward integrating past approaches of data collection with new approaches while incorporating the rigor associated with psychometric measurement/the science of human performance measurement. In academic or professional settings where datasets are manageable, the functions performed by the automated tools are performed by various types of analysts. However, in settings such as military contexts where data sets are vast, continuously growing and derived from highly dynamic contexts, human analysts would not be practical nor sufficient for performing these very important functions. Thus, leveraging current advances in computing and software technology to automate these functions is of huge value to the warfighter.

Again, it is important to note that this paper should be seen as a starting point for an intellectual discourse between academics and practitioners about balancing the science while advancing the technology regarding human performance measurement. As the science advances so should the technology and vice versa.

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

Anastasi, A., Urbina, S. (1997). Psychological testing (7th ed.). Upper Saddle River, NJ, US: Prentice Hall/Pearson Education.

Atkinson, B. F., Tindall, M., Killilea., J., Tolland, M., & Dean, C. (2017). Standardizing human performance measurement for ease of data analytics. *51st Interservice/Industry Training, Simulation and Education Conference (I/ITSEC), Orlando, FL*.

Balcazar, F., Hopkins, B., & Saurez, Y. (1985). A critical, objective review of performance feedback. *Journal of Organizational Behavior Management*, 7, 65-89.

Bartram, D. (2005). The great eight competencies: a criterion-centric approach to validation. *Journal of Applied Pychology*, 90, 1185-203.

Bernardin, H., & Pence, E. (1980). Effects of rater training: Creating new response sets and decreasing accuracy. *Journal of Applied Psychology*, 65, 60.

Bobko, P., & Coella, A. (1994). Employee reactions to performance standards: A review and research propositions. *Personnel Psycholog*, 47, 1-29.

Borman, W. C. (1975). Effects of instructions to avoid halo error on reliability and validity of performance evaluation ratings. *Journal of applied psychology*, *60*(5), 556.

Cronbach, L. (1951). Coefficient alpha and the internal structure of tests. *Psychometrika*, 16, 297-334.

Dalessio, A. (1998). Using multisource feedback for employee development and personal decisions. *Performance appraisal: State of the art in practice.* 278-330.

Dean, J. (2014). *Big data, data mining, and machine learning: value creation for business leaders and practitioners*. John Wiley & Sons.

Earley, P. C., Northcraft, G. B., Lee, C., & Lituchy, T. R. (1990). Impact of process and outcome feedback on the relation of goal setting to task performance. *Academy of Management Journal*, *33*(1), 87-105.

Elliott, S. N., & Bretzing, B. H. (1980). Using and updating local norms. *Psychology in the Schools*, *17*(2), 196-201.

Fischer, K., & Jungermann, H. (1996). Rarely occurring headaches and rarely occurring blindness: Is rarely= rarely? The meaning of verbal frequentistic labels in specific medical contexts. *Journal of Behavioral Decision Making*, *9*(3), 153-172.

Franic, D. M., & Pathak, D. S. (2000). Communicating the frequency of adverse drug reactions to female patients. *Drug Information Journal*, *34*(1), 251-272.

Guion, G. (1998). The role of perception in the sound change of velar palatazition. *Phonetica*, 55, 18-52.

Hauenstein, N. M. A. (1998). Faking personality tests and selection: Does it matter. In *MA McDaniel & AF Snell (Chairs), Applicant faking with non-cognitive tests: Problems and solutions. Symposium conducted at the annual meeting of the Society for Industrial and Organizational Psychology, Nashville, TN*.

Kelley, T. (1925). The applicability of the Spearman-Brown formula for the measurement of reliability. *Journal of Educatinoal Psychology*, 16, 300.

Latham, G., Mitchell, T., & Dossett, D. (1978). Importance of participative goal setting and anticipated rewards on goal difficulty and job performance. *Journal of Applied Psychology*, 63, 163.

MacKenzie, S. B. (2003). The dangers of poor construct conceptualization. *Journal of the Academy of Marketing Science*, *31*(3), 323-326.

Pearsonm K. (1907). On further methods of determining correlation. *Dulau and Company*.

Pritchard, R. (1990). Measuring and improving organizational productivity: A practical guide. *Greenwood Publishing Group*

Pritchard, R. D., Harrell, M. M., DiazGranados, D., & Guzman, M. J. (2008). The productivity measurement and enhancement system: a meta-analysis. *Journal of Applied Psychology*, *93*, 540.

Schaller, R. R. (1997). Moore's law: past, present and future. *IEEE spectrum*, *34*(6), 52-59.

U.S. Navy & American Navy Recruiting. (n.d.). Retrieved from https://www.navy.com/

Viswesvaran, C., Schmidt, F., & Ones, D. (2005). Is there a general factor in ratings of job performance? A meta-analytic framework for disentangling substantive and error influence. *Journal of Applied Psychology*, 90, 108.

Weber, E. U., & Hilton, D. J. (1990). Contextual effects in the interpretations of probability words: Perceived base rate and severity of events. *Journal of Experimental Psychology: Human Perception and Performance*, *16*(4), 781.

Woher, D., & Huffcut, A. (1994). Rater training for performance appraisal: A quantitative review. *Journal of Occupational and Organizational Psychology*, 67, 189-205.