Developing Systems-based Performance Measures: A Rational Approach

Kara L. Orvis
Pacific Science & Engineering
San Diego, CA
karaorvis@pacific-science.com

Andrew Duchon
Aptima, Inc.
Washington, DC
aduchon@aptima.com

Arwen DeCostanza
U.S. Army Research Institute of the Behavioral and Social Sciences
Aberdeen, MD
Arwen.h.decostanza@us.army.mil

ABSTRACT

Organizations, including the military, are producing an exponentially growing amount of data due to the rising use of technology including multimedia, social media, smartphones, and data sensors, not to mention conventional methods of communication such as email, chat, and phone. Strategic use of so-called big data is one way for the United States military to maintain a competitive advantage. Although the amount of potential data available is vast, there are many challenges to capturing the full potential of big data-driven strategies, including storage, curation, visualization, and analysis. Further, most approaches to utilizing big data make the mistake of relying solely on a data-driven approach, largely ignoring theory, science, and contextual expertise (Graham, 2012). The purpose of this paper is twofold: first, to present the perceived benefits of utilizing big data to monitor and measure individual and team performance in command and control environments; and second, to describe a new method of developing organizational performance measures which utilize big data but are also rooted in an understanding of theory and context. The Rational Approach to Developing Systems-based Measures (RADSM) is a series of six steps which consider theory, data availability, and a range of multi-disciplinary data analysis methods combined to develop measures of a construct of interest in a top-down and bottom-up approach. The RADSM process emerged from several years of research focused on developing unobtrusive performance measures for large military organizations. The underlying logic of RADSM is that as organizational members interact with one another, their actions – such as communicating with one another, making collaborative decisions, and executing their individual position-related duties – leave behind "trace data" that provide insight into individual and collective states. The authors of this paper have found RADSM to be a flexible approach by which to thoughtfully organize the mining of big data for performance assessment.

ABOUT THE AUTHORS

Kara Orvis is an Industrial-Organizational Psychologist and Director of Training at Pacific Science and Engineering (PSE). Dr. Orvis has expertise in the areas of leadership, team collaboration, culture, dispersed leadership, and training technologies. At PSE, she leads projects related to military assessment, formation, training, and development. Prior to joining PSE, Dr. Orvis worked as a principle scientist at Aptima and as a Post-Doctoral Research Fellow for the Consortium of Universities at the Army Research Institute where she led research projects involving teams, leadership, and training technologies. During her graduate training, Dr. Orvis helped manage an Air Force funded lab investigating process training, performance measurement, and feedback practices for multi-team systems. Dr. Orvis has also worked as a private consultant in international training. Dr. Orvis holds an M. A. and a Ph.D. in Industrial-Organizational Psychology from George Mason University and a B.A. in Psychology from Ohio Wesleyan University. She is a member of the American Psychological Association and the Society for Industrial and Organizational Psychology.

Andrew Duchon is the Capability Lead of the Organizational Performance Assessment team in Aptima's Performance Assessment and Augmentation Division. He is a project manager, researcher, and software engineer applying natural language processing to projects requiring the analysis of text to understand human behavior in a variety of domains. . He has held positions as a researcher at the Advanced Telecommunications Research Institute in Japan, an NSF postdoctoral fellow at Brown University, the Director of Technology at Simpli.com (an internet search company) and a private consultant in the finance industry. Dr. Duchon holds a Ph.D. in Cognitive Science from Brown University, and a B.A. in Psychology and the Integrated Science Program from Northwestern University. He is a member of the Association for Computational Linguistics.

Arwen DeCostanza Dr. Arwen DeCostanza is a Senior Research Psychologist at the U.S. Army Research Institute. Dr. Decostanza has expertise in the areas of team and organizational performance, unit cohesion, interpersonal trust, culture, and personnel selection. Her research focuses on enhancing performance in complex military teams. At ARI, Dr. DeCostanza leads research concentrating on organizational effectiveness in distributed command and control environments and the development of a system for unobtrusively tracking unit cohesion over time, examining theoretical, methodological, and system development issues. Dr. DeCostanza received her Ph.D. in Industrial and Organizational Psychology from The George Washington University and her B.A. in Psychology from Wake Forest University. She is an active member of SIOP and APA and serves as webmaster for APA's Division 19 (Military Psychology).

Developing Systems-based Performance Measures: A Rational Approach

Kara L. Orvis
Pacific Science & Engineering
San Diego, CA
karaorvis@pacific-science.com

Andrew Duchon
Aptima, Inc.
Woburn, MA
aduchon@aptima.com

Arwen DeCostanza
U.S. Army Research Institute of the Behavioral and Social Sciences
Aberdeen, MD
Arwen.h.decostanza @us.mail.mil

INTRODUCTION

Organizations, including the military, are producing an exponentially growing amount of data due to the rising use of technology including multimedia, social media, smartphones, and data sensors. In addition, military organizations are using a multitude of direct communications systems including e-mail, chat, radio, phone, VoIP, and loudspeaker. Military staffs are also communicating via document sharing on an assortment of systems (e.g. Army Battle Command Systems). Finally, there are emerging technologies (Cattuto et al., 2010; Olguín et al., 2009) which can capture data regarding face-to-face interactions. All of these systems are producing enormous amounts of data. In an effort to maintain a competitive advantage, the US could leverage this data to better understand how the individual and groups within organizations are functioning in order to set the foundation for improving performance.

The purpose of this paper is twofold: first, to present the benefits of utilizing big data to monitor and measure individual and team performance in command and control environments; and second, to describe a new framework for developing organizational performance measures utilizing big data, referred to as Rational Approach to Developing Systems-based Measures (RADSM).

Using Big Data for C2 Performance Measurement

The environments in which military staff organizations operate have grown increasingly complex. The number of people and tools that allow for collaboration across time and space boundaries continue to evolve and grow. In order to be successful, teams should be able to combine the constant inflow of information about the changing environment with information about their own collaborative performance in order to quickly and effectively adapt their behavior to dynamic situations.

Traditionally, assessments for individuals and groups have relied on self-ratings, the opinions of instructors, subject-matter experts, or outcome measures associated with quantifiable objectives (e.g., percentage of targets hit). However, these traditional measurement approaches are limited by things such as cost of deployment, human biases, and delay in results. Further, these measures overlook the biggest advantage of networked collaborations – the largely untapped plethora of member-generated data available within the collaboration tools and systems. Although data from these team environments can be challenging to access, with social and ethical concerns that need to be addressed when using such data in the workplace (McCormack, Duchon, Geyer, & Orvis, 2009), a wealth of useful information could be derived from analyzing the content and flow of communications, the spatially and temporally varying social networks, and human-computer interactions. Additional information such as organizational structure, task objectives, and the members' backgrounds may also be available to provide greater context for these analyses.

System-based measures offer a number of advantages over self-report and observer ratings. First, system-based measures are extremely efficient; they can reliably collect a wealth of information about team processes and performance, and are not subject to the limitations of trained observers (e.g., fatigue, distraction, missed observations). Second, system-based measures tend to be well-received by the participants, because the results are perceived as "objective"; they are not subject to rating errors, such as the "halo effect" (Murphy & Balzer, 1989).

Finally, and perhaps most importantly in a military setting, system-based measures can be captured unobtrusively, requiring no time or effort from participants or observers.

System-based measures have been used in the past for measuring organizations and teams. However this has largely occurred in research environments, with the majority of assessments focusing on overall performance (e.g. wins and losses). Few attempts have been made at using systems-based data to assess team and organizational states, such as group trust or cohesion, especially in large organizations; for these aspects, measurement approaches have leaned heavily on validated questionnaires. By focusing on team states, rather than outcome measures, measures which generalize across environments and may be more likely. While the type of task no doubt influences what is required for a team to perform well, more trust and cohesion are likely to improve results independent of the task. Thus, the development of unobtrusive methods for measuring team states could have application to a wide variety of domains.

Further, employment of methods such as network analysis or automated language analysis on team and organization communications is rare, particularly to assess team constructs such as trust, cohesion, and shared mental models (though see Carley, 1997). The intent of this research was to revisit the use of systems-based data in measuring team states. It is the authors' opinion that systems-based measurement approaches have been limited by the platforms used and the imagination and method by which measures were constructed. It seems vital that science push forward methods by which organizations can meaningfully measure team states and processes, predict performance, and potentially augment and intervene to improve performance.

RADSM APPROACH TO SYSTEM-BASED MEASURES

As part of a larger research program to develop unobtrusive measures of organizational performance, the research team derived a method to develop systems-based measures of psychological constructs referred to as the Rational Approach to Developing Systems-based Measures (RADSM). The RADSM process is dependent on the underlying logic that as team members interact with one another, their actions – such as communicating with one another, making team decisions, and executing their individual position-related duties – leave behind "trace data" that can be reconstructed, either in real-time or post-hoc, to understand an individual or the team's states and processes.

The use of trace data is not a new concept, as researchers have long declared that individuals are the sum total of their behavior and experience (Allport, 1937) and that when a group of people are engaged in a task that their communication reveals the group's cognitive processes (Cooke, Gorman, & Rowe, 2009; Cooke, Gorman, & Winner, 2007). Although the methods of data collection, analysis, and synthesis presented in this research may be different from traditional psychological measurement approaches, conceptually the approaches are similar, if not identical to, other tried and true approaches to measurement development in psychology.

Parallels to Survey Measure Development

Traditional, survey-based measures ask individuals to report their perceptions of aspects of themselves, specific to a construct of interest. One particular type of survey-based measure is biodata. The premise behind biodata measures is that past patterns of an individual's behaviors in particular situations are predictive of future behavior and even qualities of their personality (Mumford & Owens, 1987, 1992). Biodata items are specified as different from other surveys measuring psychological constructs in that they focus on observable factors and are therefore more objective in nature (Mumford & Owens, 1987). Biodata measures have been successfully developed for assessing attitudinal and cognitive constructs, and have been used in practical settings (e.g. Stokes, Mumford, & Owens, 1994). Given these factors, as well as the focus on observable and verifiable data, our work models the RADSM approach after the biodata approach.

Although we model the biodata approach to measure development, distinguishing measures derived through the RADSM method and traditional biodata measures is important. First, within RADSM, information does not come from self-reports but rather by collecting information that is available in various systems with which a person interacts. To illustrate this difference, we extend an example provided by Farmer (2007) that distinguishes biodata measures from typical self-report personality measures of extroversion (see Table 1). As depicted in the table, RADSM items are almost identical in nature to biodata items, but the qualities of the measures are distinct. Second, the RADSM approach incorporates more advanced statistical analyses. For example, identifying the extent to which a person is engaging in activities may require the application of social network or language analysis methods to mine available data. Third, biodata measurement approaches have been predominantly used for individual, and not team, level measurement. However, we believe the concept is applicable to teams as well. In summary, the measures derived through the RADSM approach are distinct from biodata measures in the following ways: 1) the data is gathered via systems information solely, without any self-report, 2) the analysis methods delve beyond mere frequency counts and into numerical weightings derived from network analysis, natural language processing, and dynamical systems, and 3) the level of measurement can be focused on the team.

Table 1: Examples Comparing Systems-Based Data Measurement Approach to Traditional Survey-Based Measurement Approaches for Individual Level of Analysis

Type of Item	Item	Qualities
Personality survey (self-report)	"I really enjoy talking to people"	Solicit information regarding an individual's predisposition or general behavioral tendency. Focus is on the person's disposition.
Biodata (self-report)	"How often do you get together with friends"	Focus on prior behavior and experiences occurring in specific situations. Capture dispositional and environmental aspects. Focus on "objective" measures such as self-reported frequency.
RADSM Approach (systems-based data)	"How often does this person get together with friends?"	Similar to biodata item but network data is collected from systems such as calendars or accepted party invitations in e-mail.

Steps to RADSM Process

Although the development of many biodata measures is not described in great depth, Mumford and Stokes (1992) describe a "rational" construct-oriented item generation approach to developing biodata measures. Consider an example of this method in Zaccaro et al. (1995) who developed the Background Data Measure of Social Intelligence. Their first step was to bring together experts in the methodology of background data as well as expertise in the construct of interest (in this case social intelligence). The expert panel was then asked to generate items related to the sub-constructs using theoretical and operational guidelines and definitions. After a complete list of items was developed, the panel then assessed each item by its face validity and susceptibility to response biases. At its core, the "rational" approach to developing biodata measures is to integrate subject matter expertise, theory, and validation.

RADSM follows a similar approach to Mumford's rational approach to biodata development, but integrates steps related to developing systems-based measures rather than self-report. After a few iterations, the research team settled on a series of six steps to conduct the RADSM measurement development process which are depicted in Figure 1. The overall approach is that theory, data availability, and a range of multi-disciplinary data analysis methods combine to develop measures of the construct of interest applying a top-down and bottom-up approach. The top-down approach refers to the process of selecting observable behaviors and attributes which are theoretically relevant to the construct of interest and specific to the context of the work environment. This is very similar to other survey-based approaches to developing items for the measurement of psychological constructs. The bottom-up approach refers the process of identifying which systems-based data and analysis methods can provide insight into those theory and context driven behaviors and attributes. The difference between Mumford's approach and the RADSM approach include: 1) the ability to monitor behaviors unobtrusively; 2) the range of systems data that can be leveraged; and 3) the methods by which one can analyze or infer meaning from systems data.

The following sections walk through the RADSM steps to develop system-based measures of constructs. While the approach we outline below is applicable to any type of system data, we focus here on organizational communications data that can be accessed unobtrusively (e.g., e-mail).

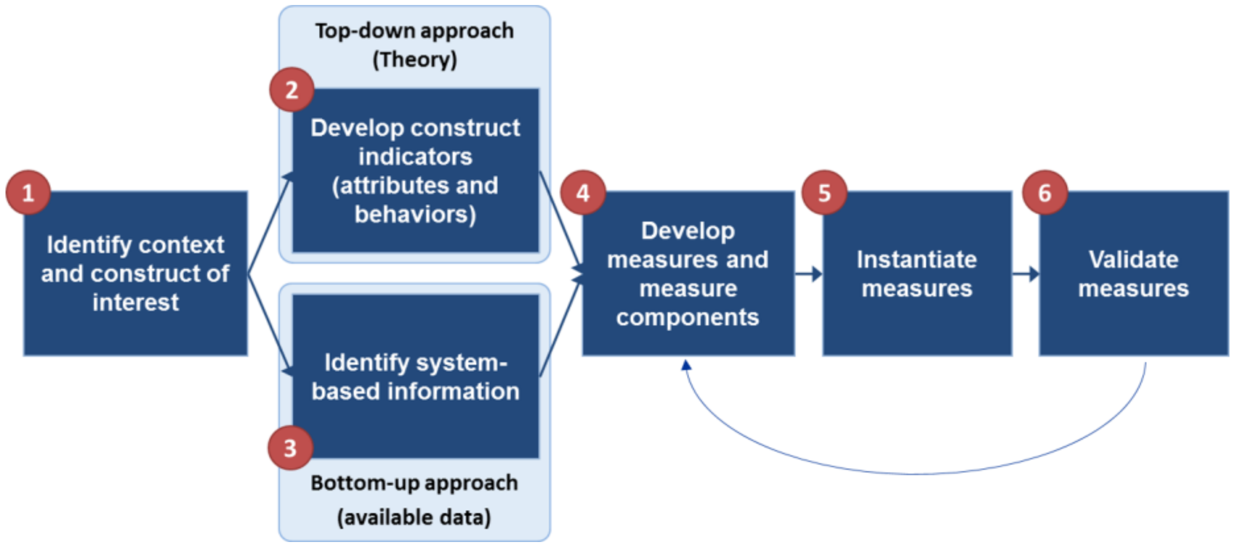

Figure 1. RADSM Steps

EXAMPLE OF DEVELOPING A RADSM MEASURE

Step 1: Identify context and measurement construct of interest

The first step of the RADSM process involves specifying the measurement constructs of interest. This step is not unique to the RADSM process and is often the first step in many other measurement development approaches. Key considerations in this step are: specifying the construct of interest, understanding conceptually what the construct is, distinguishing the construct of choice from other related constructs, and linking the construct in a meaningful way to the context in which it will be observed.

Shared Situational Awareness
To provide a running example, we will use *shared situational awareness* (SSA) as the targeted construct. Endsley & Jones (1997) define SSA as the extent to which "team members possess the same SA on shared SA requirements" (Endsley & Jones, 1997, p 47; 2001, p.48). SSA focuses on the overlap in knowledge requirements for all team members and is critical to the interdependency of the team members. It is important that team members are able to share the information that is relevant to their interdependent coordination, not overloading other team members with ancillary information related to an individual team member's job. This is particularly true for the military. For the purpose of this example, we will focus on developing a systems-based measure for SSA. Additionally, the context will focus on Army Battalion and Brigade staff environments.

Step 2: Develop a list of attributes and behaviors of the target construct within the context being studied

The second step of the RADSM process is to develop a list of attributes and behaviors of the target construct within the context being studied. These are referred to as *construct indicators* which basically means they are observable behaviors or attributes that can provide some insight into the construct of interest within the targeted environment. It is imperative to have experts in the construct and context of interest to participate in this step. Having such experts will ensure the construct is well represented and the specific idiosyncrasies of the context are included in that representation. Example indicators and their theoretical and contextual underpinnings are in Table 2 for SSA.

Table 2: Shared Situation Awareness and Example Indicators

Shared SA Description	Example Construct Indicators
One subcomponent of SSA, devices, as defined by Endsley & Jones (1997), references how teams use the devices they have available for sharing information, which can include direct communication (both verbal	• Team members with high SSA will send and receive messages which mention the devices by which they share information (positive indicator).

and non-verbal), shared displays (e.g., visual or audio displays, or tactile devices), or a shared environment. Within an Army operations center, SSA is dependent on team members exchanging mission critical information with one through these media.	
Another subcomponent of SSA, requirements, as defined by Endsley & Jones (1997), references the extent to which team members know what information need to be shared with their team.	• Teams with high SSA have a good understanding of individual team member tasks and provide unique, non-redundant, information using their information sharing devices; therefore more dissimilarity in their messages will reflect better SSA.

Step 3: Identify system-based information

The purpose of this step is to assess what systems-based information may be available within the context being observed. This includes developing a thorough list of types of data available (e.g., e-mail communication, text, resume) and general methods of analysis (e.g., language analysis, counts, network analysis). Other aspects of the data may be considered, such as data gathering sensitivity or security issues. The primary goal of this task is to exhaust all potential sources of meaningful data without being bounded by the specific types of information required.

When analyzing the potential data available, it is helpful to think about what data will tell you about the individuals, the groups they work in, and the tasks they are working on. Resources such as promotion packages, resumes, and biographies provide textual descriptions of individual professional skills, educational background, work history, and work assignments. This type of information is vital when one needs to identify and assess information about individuals or to identify their groups. Resources such as e-mail, phone, and instant messaging logs provide insight into the communication network among individuals and can be used to assess teamwork processes and states. Timecards and login sheets provide some assessment of time spent on tasks and with other members of a group. Table 3 provides a range of systems data we identified as available to most Army staff organizations at the Brigade and Battalion echelons. In addition, in this research we have explored the use of Sociometric Badges (Olguín et al., 2009) which are small devices worn around the neck which can record physical interactions. As such technology becomes miniaturized, and if our work shows its usefulness, such data may also be common in the future as this technology could be integrated into standard identification cards, uniforms, etc. This list is not exhaustive and is meant to serve as an example of data that may be useful in constructing measures.

Table 3: Example Data Resource Inventory for Generic Command Post Environments

Data Type	Example Data Source	Information Obtained
Face-to-face interaction	Sociometric Badge: IR sensor	Who is facing whom when
Physical Proximity	Sociometric Badge: Bluetooth radio sensor	Who is near whom when
Face-to-face vocal communication	Sociometric Badge: Audio sensor	Who is speaking to whom when and in what tone
Individual physical movement	Sociometric Badge: accelerometer	Who is individually active and what groups are “in sync” when
Email	Microsoft Outlook	Who interacted with whom about what when and how did they talk about it
Text chat	LeafChat	Who interacted with those present within a chat room about what when and how did they talk about it
Meeting space records	Adobe Connect video	Who interacted with those present in larger group meetings when

VoIP call logs	Cisco Call Manager	Who interacted with whom when
Radio	Joint Virtual Tactical Radio (JVTR)	When a radio band was active and who monitors that band
Operational collaboration	Command Post of the Future (CPoF)	Who created and accessed which objects in the common operating picture (COP) tool when
File Systems	Microsoft SharePoint	Who created and accessed which documents concerning what when

As evidenced by Table 3, in any organization, there is a wide array of systems information that can be utilized for analysis. There are many ways in which that data can be analyzed and processed, from simple counts and averages to the integration of multiple data sources and complex analysis methods such as text analysis and network analysis. A measure of a participant's shared situation awareness with others on his/her team suggests the participant should be aware of the "things" being discussed by his/her teammates—these "things" would be present in the content of communications. Text analysis can get at this content. Shared SA is also somewhat dependent on the network by which information is shared. Network analysis can provide some insight into how information is being passed. For this example, we focused on one methods of analysis to construct a measure of SSA; language analysis.

Step 4: Formalize definitions of measures and measurement components

The purpose of this step is to match the "indicators" identified in Step 2 to the data and analyses available identified in Step 3. Basically this means looking for what systems information within a particular context can provide an objective assessment of the indicator. The overall formula for that is called an "item" and is meant to mirror survey-based items. A "measure" is then compiled from the appropriate assortment of all "items" within the given construct of interest (refer to Figure 2).

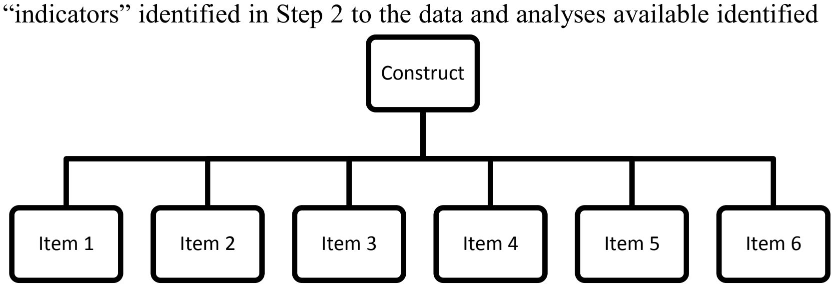

Figure 2. Multiple Item Approach to Measure Development

Note that this approach is similar to constructing survey-based measures. However, aggregating these items to complete a measure may look very different when dealing with systems-based items. In the simplest example, aggregating the items can be similar to a survey based measure in that all items are treated equally and a normalized average of all items can serve as the construct "measure." As another example, a pure systems-based measure might compile the items for SSA from, say, email and chat and weight them according to their individual reliability. Finally, the items may show compensatory relationships which could also be built into a "measure" algorithm. Understanding best methods for aggregating systems-based items warrants attention but is beyond the scope of this paper. For the current effort, we concentrate on understanding the value of individual items as building blocks.

Example Items

Table 4 provides two example items which could be considered as part of a measure assessing component of SSA. The paragraphs below describe the construction of each item.

Item #	Item	Analysis Methods	Data Sources
1	**Systems Messages:** The degree of messages a person is receiving which reference the systems they are using for information sharing (positive indicator)	Language Analysis (Dialogue Act Analysis)	E-mail
2	**Topic Dissimilarity:** The degree to which individuals within a team are sending and receiving dissimilar information.	Language Analysis (Topic Modeling)	E-mail

Item 1. As part of Step 2, we identified that within an Army operations center, SSA is dependent on team members exchanging mission critical information with one another using the devices meant for information sharing. Participants who are receiving information through the information channels in their work setting should have higher levels of SSA. In a Division staff environment teams are commonly sharing information using e-mail, phone, and face-to-face interactions. Individuals with high levels of SA are likely to mention the devices they are using for information sharing. Using text analysis, we can analyze e-mail communications to assess the degree to which individuals are receiving e-mails from their team members which reference the devices they are using for communication. The extent to which individuals are receiving messages which include content regarding their information sharing devices are more likely to have higher SSA than those who are not.

Item 2. The second item notes the extent to which individuals in a team are discussion dissimilar information. In command staff teams with high SSA, it is likely that individuals will have a good understanding of individual team member tasks and will provide unique, non-redundant information. This can be assessed through topic modeling.

Calculating a Team Level Construct

To make meaningful inferences about the objective data, it is important to consider how to aggregate it to capture the multilevel nature of SSA. The items represented in Table 4 do not yet represent SSA, which is a team level construct. They may represent an individual's SSA within a team but not the overall SSA construct as a team measure. To truly form a SSA level equation all indicators should be aggregated across team members. By doing so, team-level SSA could be inferred.

There are many options in deciding how to aggregate these example individual level items to represent a team level construct. Three examples include the following:

- Option 1: Total count across all team members. This provides an overall score for the team but may not offer insight into variability across team members (i.e. if members are left out).
- Option 2: Lowest calculation for one team member. Looking at the lowest scoring individual may provide some insight into the team.
- Option 3: Calculation considers total variance of the team and variance between team member pairs. This calculation would provide insight into all the exchanges across all pairs of team members.

Step 5: Instantiate measures

The purpose of step five is to reach into the data and pull out the items developed in step 4. In order to instantiate systems-based measures, one needs to collect, store, correlate and manage the vast amounts of data. Through our research we have found that most military operational, training, research facilities are not collecting or storing these data in a way that facilitates any type of real-time or post hoc measurement approach. For this work, Aptima developed a suite of tools to collect, store, and analyze communications data in order to instantiate these measures. The purpose of this paper is not to highlight this technology but we felt it was important to mention the difficulties with instantiating these measures.

Step 6: Validate Measures

The final step of any measurement development is providing evidence of validity. A measure is considered valid when it measures what it intends to measure or the degree to which it accomplishes what it is supposed to accomplish (Pedhauser and Schmelkin, 1991). There are various types of validity evidence including content validity, criterion-related validity, and construct validity.

For this research, we have been exploring the construct validity of the items. Construct validity considers whether a scale measures the construct that it intends to measure. There are several ways one can assess construct validity. The first is through an assessment of *face validity* by subject matter experts (e.g. Anastasi, 1988). A measure can be assessed through experts weighing in on whether they think the measure looks like what it is intended to measure. This form of validation is commonly used when it is not possible to assess construct validity through other means. A more preferred method of assessing construct validity is by providing evidence of convergent and divergent validity (Campbell & Fiske, 1959). Convergent validity can be shown by statistically testing whether the measure correlates with measures meant to assess the same construct or highly related constructs. Discriminate validity is shown by providing evidence of the opposite: that the measure of interest does not correlate with other unrelated constructs. At a minimum, the RADSM approach requires an assessment of face-validity through subject matter expert review. For this paper, we will highlight example research assessing the convergent validity of the two items constructed above. It should be noted that no measure should consist of only two items. In the overall research program we are validating hundreds of items for several constructs across many exercises.

Participants

Data was obtained from a Brigade Warfighter Exercise (BWFX) conducted by the Mission Command Training Program. The exercise focused on operations training for a National Guard Infantry Brigade Combat Team, which included Soldiers from the Brigade and supporting Battalion staffs. The exercise lasted 10 days with the scenario and staff running 24 hours a day. Hundreds of personnel were involved in the exercise including the unit Soldiers, exercise operations, exercise control, and analysis. This validation effort focused on a sample of Soldiers within the unit (n=55). The sample size was driven by constraints associated with the number of sociometric badges that were available for the study. Further, the study environment required the use of written surveys which had to be handed out and collected across a large physical landscape within a 24 hour period. The selected participants comprised key roles within the Brigade and Battalion units. Further, they could be configured into 13 meaningful teams. The brigade teams represented a subset of functional cells as outlined by ATTP 5-01.1. Participant teams for each of the six battalion units were comprised of the battalion commander, the intelligence officer (S2), the operations officer (S3), and the executive officer (XO).

Data Collection Procedure

There were two methods of measures collected throughout this exercise, systems-based and survey-based. For the systems-based measures, the raw data (e.g. e-mail) was collected using tailored software and processed in a secure facility at a later date. For the survey data, three surveys were administered over the course of the exercise. The first survey gathered demographic data such as familiarity with the unit, prior staff experience, rank, and position within the exercise. Administration of the first survey began during the initial exercise meeting with the unit (prior to the exercise starting) along with the informed consent forms. A researcher addressed the participants, describing the research and informing the participants that the research was voluntary. About 80% of the selected participants were present. About 60% of the participants then filled out Survey 1. Within a 24 hour period, the researchers located and approached the participants that were not at the initial meeting and requested their consent. The researchers also collected the remaining 40% of the missing surveys. At the end of that 24 hour period, fifty-five of the participants agreed to participate in the survey administration. Fifty of those same participants agreed to also participate in the sociometric badge data collection. The second survey was given after completion of the Military Decision Making Process, or planning period, which lasted approximately three days. The surveys were handed out with the sociometric badges and collected within 36 hours, with the majority of surveys collected within a 24 hour window. The third survey was handed out well into the mission execution phase of the exercise. These were also handed out with the sociometric badges and collected within 36 hours. The average response rate for all survey administrations was approximately 90%.

Measures

SSA was measured using an abridged version of the Shared SA Inventory (SSAI) which has been administered in other staff exercise settings such as OmniFusion 2010 (Scielzo, Strater, Tinsley, Ungvarsky & Endsley, 2009). The SSAI assessed operator's subjective appraisal of shared SA at the time in which it is administered. The abridged 24-item SSAI was designed as a self-report measure to assess the four Shared SA subcomponents: team devices (the apparatus and electronic interfaces used by individuals and teams to acquire, process, and disseminate information), team mechanisms (indicating common training and shared mental models), team procedures (including contingency planning and information sharing strategies), and team requirements (SA information needs). This measure indicated acceptable reliability (Cronbach's Alpha=.87).

Data Aggregation

Both the survey and systems-based measures were aggregated over the exercise planning phase (Time 1) during which the units engaged in the Military Decision Making Process (MDMP) and the mission execution phase (Time 2). Given the small sample size, stacking the data across these times improved the power to some degree.

Results

Correlations were run between the three systems-based items and the survey-based SSA measure to explore the validity of those items. The correlations indicate that both the *systems messages* item and the topic *dissimilarity item* were significantly correlated ($p<.05$) with the survey-based measure of SSA, providing some evidence of convergent validity.

Limitations of Validation Approach

While it makes sense to validate the objective indicators of a construct against a self-report or observer rating of the same construct, one must consider whether this is an 'apples to oranges' comparison. To the extent that the objective measures are intended to represent the construct, then no major significance should be placed on whether or not the objective measures relate to the survey measures given possible definitional and measurement limitations. If the objective measures are not correlated with the survey measures, it could indeed suggest that they are measuring different constructs. However, it would be impossible to infer whether the survey-based or continuous data more closely represents the construct of interest. Further research is required to determine whether the objective continuous data is a viable or, perhaps, more accurate alternative for inferring team states. To properly validate the objective measures of a construct against another measure of the same construct (whether it be a survey or observer –based measure); a few best practices should be followed. First, the level of analysis (individual vs. team) should be the same across the two measures (comparing team-level to team-level, or individual-level to individual-level). Comparing individual ratings on one measure to team ratings on the other measure would lead to faulty inferences. Additionally, the level of molarity at which a construct is operationalized (i.e., item-, component, construct) should be consistent across measures as well. For example, if the self-report items are averaged to form an overall score, then that score should be compared to an overall score on the objective measure, rather than an individual indicator. By following these best practices, the validation effort will provide the most appropriate information regarding the extent to which the two measures relate to one another.

DISCUSSION

The development of unobtrusive measures using big data can be of great benefit to large organizations. Currently most practitioners rely on data-driven approaches. The authors of the paper have found the RADSM process to be a flexible tool by which to organize the mining of big data which is supported by theory and contextual expertise as well as data. Employment of the RADSM process requires a great deal of creative thought from both construct and context experts. Our intentions are that it serves as a useful process for others who wish to use big data for measurement purposes.

ACKNOWLEDGEMENTS

Work presented in this paper was conducted under contract *W91WAW-09-C-0058* from the U.S. Army Research Institute for the Behavioral and Social Sciences (ARI) awarded to Aptima, Inc. We would also like to acknowledge and thank the Mission Command Training Program for their support.

REFERENCES

Allport, G. W. (1937). *Personality: A psychological interpretation*. New York: Holt, Rinehart, & Winston.

Anastasi, A. (1988). *Psychological testing.* New York, NY: Macmillan.

Campbell, D. T. and Fiske, D.W. (1959). *Convergent and discriminant validation by the multitrait-multimethod matrix.* (56). Psychological Bulletin. pp. 81–105.

Carley, K. M. (1997). Extracting team mental models through textual analysis. Journal of Organizational Behavior, 18(S1), 533–558.

Cattuto, C., Van den Broeck, W., Barrat, A., Colizza, V., Pinton, J.-F., & Vespignani, A. (2010). Dynamics of person-to-person interactions from distributed RFID sensor networks. *PloS one*, *5*(7), e11596.

Cooke, N. J., Gorman, J. C., & Rowe, L. J. (2009). An ecological perspective on team cognition. In E. Salas, G. F. Goodwin, & C. S. Burke (Eds.), Team effectiveness in complex organizations: 157-182. New York: Routledge, Taylor & Francis Group.

Cooke, N. J., Gorman, J. C., & Winner, J. L. (2007). Team cognition. In *Handbook of Applied Cognition* (Vol. 2, pp. 239–268).

Endsley, M.R. & Jones, W.M. (1997). Situation awareness, information dominance, and information warfare (No. AL/CF-TR-1997-0156). Wright-Patterson AFB, OH: United States Air Force Armstrong Laboratory.

Farmer, W. L. (2007). A Brief Review of Biodata History, Research, and Applications. Millington, TN: Navy Personnel Research Studies and Technology.

Graham, 2012. Big data and the end of theory? The Guardian. Retrieved http://www.guardian.co.uk/news/datablog/2012/mar/09/big-data-theory on 28 June 2013.

McCormack, R. K., Duchon, A., Geyer, A., & Orvis, K. (2009). A Case Study in Data Mining for Automated Building of Teams. In Social implications of data mining and information privacy: Interdisciplinary frameworks and solutions (pp. 249–267). Hershey, PA: IGI Global.

Mumford, M. D. & Owens, W. A. (1987). Methodology review: Principles, procedures, and findings in the application of background data measures. Applied Psychological Measurement, 2, 1-31.

Mumford, M. D., & Stokes, G. S. (1992). Developmental determinants of individual action: Theory and practice in applying background measures. In M. D. Dunnette & L. M. Hough (Eds.), Handbook of industrial and organizational psychology: Vol. 3 (2nd ed., pp. 61-138). Palo Alto CA: Consulting Psychologists Press, Inc.

Murphy, K.R., & Balzer, W.K. (1989). Rater errors and rating accuracy. *Journal of Applied Psychology, 74*(4), 629-624.

Olguín, D. O., Waber, B. N., Kim, T., Mohan, A., Ara, K., & Pentland, A. (2009). Sensible Organizations: Technology and Methodology for Automatically Measuring Organizational Behavior. *IEEE Transactions on Systems, Man, and Cybernetics, Part B (Cybernetics)*, *39*(1), 43–55. doi:10.1109/TSMCB.2008.2006638

Pedhazur, E. J., & Schmelkin, L. P. (1991). Measurement, design, and analysis: An integrated approach. Hillsdale, NJ: Erlbaum.

Scielzo, S., Strater, L., Tinsley, M. L., Ungvarsky, D., & Endsley, M. (2009). Developing a Subjective Shared Situation Awareness for Teams, *Human Factors and Ergonomics Society 53rd Annual Meeting*. San Antonio, TX.

Zaccaro S. J., Gilbert, J. A., Zazanis, M. M., & Diana, M. (1995). Investigating a background data measure of social intelligence. (Technical Report, No. 1024). Arlington, VA: U.S. Army Research Institute for the Behavioral and Social Sciences.