NASA/TM–20210021987

Standard Measures for Use in Analog Studies, ISS, and Research for Long-Duration Exploration Missions

Elizabeth M. Wenzel
NASA Ames Research Center

August 2021

NASA STI Program…in Profile

Since its founding, NASA has been dedicated to the advancement of aeronautics and space science. The NASA scientific and technical information (STI) program plays a key part in helping NASA maintain this important role.

The NASA STI program operates under the auspices of the Agency Chief Information Officer. It collects, organizes, provides for archiving, and disseminates NASA's STI. The NASA STI program provides access to the NTRS Registered and its public interface, the NASA Technical Reports Server, thus providing one of the largest collections of aeronautical and space science STI in the world. Results are published in both non-NASA channels and by NASA in the NASA STI Report Series, which includes the following report types:

- TECHNICAL PUBLICATION. Reports of completed research or a major significant phase of research that present the results of NASA programs and include extensive data or theoretical analysis. Includes compilations of significant scientific and technical data and information deemed to be of continuing reference value. NASA counterpart of peer-reviewed formal professional papers but has less stringent limitations on manuscript length and extent of graphic presentations.

- TECHNICAL MEMORANDUM. Scientific and technical findings that are preliminary or of specialized interest, e.g., quick release reports, working papers, and bibliographies that contain minimal annotation. Does not contain extensive analysis.

- CONTRACTOR REPORT. Scientific and technical findings by NASA-sponsored contractors and grantees.

- CONFERENCE PUBLICATION. Collected papers from scientific and technical conferences, symposia, seminars, or other meetings sponsored or co-sponsored by NASA.

- SPECIAL PUBLICATION. Scientific, technical, or historical information from NASA programs, projects, and missions, often concerned with subjects having substantial public interest.

- TECHNICAL TRANSLATION. English-language translations of foreign scientific and technical material pertinent to NASA's mission.

Specialized services also include creating custom thesauri, building customized databases, and organizing and publishing research results.

For more information about the NASA STI program, see the following:

- Access the NASA STI program home page at http://www.sti.nasa.gov

- E-mail your question via to help@sti.nasa.gov

- Phone the NASA STI Help Desk at (757) 864-9658

- Write to:
  NASA STI Information Desk
  Mail Stop 148
  NASA Langley Research Center
  Hampton, VA 23681-2199

NASA/TM—20210021987

Standard Measures for Use in Analog Studies, ISS, and Research for Long-Duration Exploration Missions

Elizabeth M. Wenzel
NASA Ames Research Center

National Aeronautics and
Space Administration

Ames Research Center
Moffett Field, California

August 2021

Available from:

NASA STI Program
STI Support Services
Mail Stop 148
NASA Langley Research Center
Hampton, VA 23681-2199

This report is also available in electronic form at http://www.sti.nasa.gov
or http://ntrs.nasa.gov/

Table of Contents

List of Figures and Tables

List of Figures

List of Tables

Acronyms and Definitions

3D	3-dimensional
AIAA	American Institute of Aeronautics and Astronautics
ANSI	American Standards Institute
app	application
AR	augmented reality
ARC	Ames Research Center (NASA)
ARED	Advanced Resistive Exercise Device
AWAS	Aircrew Workload Assessment System
CO_2	carbon dioxide
COBRA	Comprehensive Oculomotor Behavioral Response Assessment
Con Ops	Concept of Operations
DoD	Department of Defense
DST	Deep Space Transport
ESA	European Space Agency
EVA	extra vehicular activity
fMRI	functional Magnetic Resonance Imaging
fNIRS	functional Near-Infrared Spectroscopy
HCAAM	Human Capabilities Assessments for Autonomous Missions
HCI	Human-Computer Interaction
HCTS	Human Computer Trust Scale
HERA	Human Exploration Research Analog
HFBP	Human Factors and Behavioral Performance
HRP	Human Research Program
HRV	heart rate variability
iQ&A	Question and Answer application for iOS devices
ISA	Instantaneous Self-Assessment
iSHORT	iOS-Based SHORT
ISS	International Space Station
JITT	just-in-time training
JSC	Johnson Space Center (NASA)
LCD	liquid crystal display
LDEM	long duration exploration mission
LISA	Linear Integration of Speed and Accuracy
MAE	mean absolute error
MCC-H	Mission Control Center-Houston (NASA)
ms	millisecond
NASA	National Aeronautics and Space Administration
NEEMO	NASA Extreme Environments Mission Operations
OBT	On Board Training
OPTIMIS	Operations Planning Timeline Integration System
PVT	Psychomotor Vigilance Task
RCS	rate correct score
RMSE	root mean square error
ROBoT	Robotic On-Board Trainer
RT	reaction time

SA .....situation awareness
SAGAT .....Situational Awareness Global Assessment Technique
SART .....Situational Awareness Rating Technique
SAT.....speed accuracy trade-off
SHFH .....Space Human Factors and Habitability
SHORT .....Space Habitability Observation Reporting Tool
SME .....subject matter expert
SPAM .....Situation Present Assessment Method
SSC .....ISS laptop
SSTMF.....Space Station Training Mockup Facility
TASS.....Trust in Automated System Scale
TBI.....traumatic brain injury
TLX.....Task Load Index
VNSCOR .....Virtual NASA Specialized Center of Research
VR.....virtual reality

Standard Measures for Use in Analog Studies, ISS, and Research for Long-Duration Exploration Missions

Elizabeth M. Wenzel[1]

This document is based on a 2019 Final Report commissioned by the Human Research Program, Space Human Factors and Habitability Element, Risk of Inadequate Human-Computer Interaction (HCI).

1. Research Methods

This report summarizes the results of an effort that had the following primary aims:

1. Identify and recommend reliable, robust human factors performance measures (e.g. workload, fatigue, situation awareness, efficiency, accuracy, trust in automation) and methods for measurement relevant for spaceflight.
2. Evaluation of utility and feasibility for use in long duration exploration missions (LDEMs).
3. Identify candidate metrics for testing and evaluation in Human Exploration Research Analog (HERA) and other analog missions.

A literature review was conducted to evaluate the state-of-the-art human factors and crew state performance measures, including their reliability, repeatability and most efficient data collection methods. Interviews were conducted with Human Factors and Behavioral Performance (HFBP) Element Scientists, discipline scientists, other National Aeronautics and Space Administration (NASA) experts, and military experts to solicit performance metrics that they recommend. General recommendations regarding the best metrics for use in spaceflight, Human Capabilities Assessments for Autonomous Missions (HCAAM), analogs, and other Human Research Program (HRP) studies are provided.

2. Background

Understanding crew state and readiness to perform will be critical for successful Artemis lunar missions, Gateway, and LDEMs to Mars. These missions will have a much higher degree of crew autonomy, and are able to utilize new types of advanced displays and controls and intelligent systems that adapt to crew state and capabilities. The HRP is seeking a core set of reliable, valid human factors performance measures, that could be used to inform these systems to adapt to crew state and capabilities, assess/track crew state over time, and help inform scheduling and task allocation decisions. This work is to identify sensitive, informative performance metrics, which are as unobtrusive as possible, and suitable for use in spaceflight, HCAAM, analogs, and other HRP studies.

[1] NASA Ames Research Center; Moffett Field, California.

It should be noted that standard measures in fields like human factors are generally not comparable to medical diagnostics. For example, biomedical standard measures like a very high blood pressure or heart rate value can be judged as problematic or dangerous in all circumstances. However, even biomedical measures, when at less than extreme values, may be subject to contextual factors and ambiguity in diagnosis or meaning; an elevated intrinsic heart rate (or lower heart rate variability [HRV]) may be due to negative factors like overall fatigue and stress or more benign, even helpful, factors like increased level of attention and effort when performing a task (e.g., see Mukherjee, Yadav, Yung, Zajdel, & Oken, 2011).

Selection and interpretation of appropriate human performance measures, in particular, depends very much on contextual factors such as the specific task, operation, or human-machine system involved. An acknowledged set of standard measures related to human performance has yet to be developed either for NASA or for analogous complex human-machine systems such as those used by the military. It may be preferable and more useful to view the goal of any "standard measures" for human factors as assessing readiness to perform rather than reliably diagnosing some specific state of a human crew member. For example, there are many factors that could affect the ability of an astronaut to perform an extra vehicular activity (EVA) task during surface operations. Such factors may include physical degradation due to long-term space travel, cognitive impairments due to factors like fatigue, sleep deprivation and stress, and external factors such as the impoverished perceptual environment of a planetary surface combined with the restrictions of an EVA suit and the limitations of relevant information displays.

A 'Standardized Measures for Space Human Factors and Habitability Workshop' was conducted at NASA Johnson Space Center (JSC) in 2016. The purpose of the workshop was to inform the development of standard measures for the former Space Human Factors and Habitability (SHFH) element. There were 39 participants and included speakers from NASA who described the current practice in collecting and using measures for SHFH, as well as speakers on standard measures in domains including aviation, the military, and disaster response. A report from this workshop suggests there was much discussion but no real consensus on either identifying or characterizing a specific set of standard measures for Space Human Factors and Habitability (SHFH) (Schreckenghost, 2016).

The *Guide to Human Performance Measurements* published as an ANSI/AIAA document in 2001 discusses a number of underlying issues in human performance measurement (section 4.1). For example: "Lack of a general theory to guide performance measurement. If such a theoretical structure existed, it would relate behavioral processes within the individual to his or her performance of the task, and that task performance to total system performance. At present, only a few relationships have been discovered—for example, the inverse relationship between speed of task performance and accuracy, or performance quality." The issues described in this guide are not much different today.

Current official human factors standard measures being utilized in NASA's space program are primarily limited to Crew Notes, written at the time the crewmember is working a procedure, and Crew Comments, i.e., post-hoc comments. Both require a great deal of manpower to transcribe, analyze and interpret. Although they provide critical information about ongoing issues important to the crew, they may not address all areas of interest, adequately address changes over time, capture crew members' tendencies to gloss over performance problems, and are certainly not suitable for real-time performance monitoring.

A number of behavioral core/standard measures concerned with human health and performance have already been adopted by NASA (see *Human Exploration Research Opportunities (HERO)* 80JSC017N0001-BPBA Appendix C, 2017). Most of these are biomedical in nature although several relate to behavioral health and performance and are in the process of being tested and validated on the ISS and in ground-based and analog studies (Basner et al, 2015; Dinges et al, 2017; 2018). These include:

- Cognition Test Battery: The Cognition Test Battery is a software-based combination of 10 brief tests, evaluating different aspects of cognitive function, such as visual object learning, memory, attention, abstraction, spatial orientation, emotion recognition, abstract reasoning, complex scanning and visual tracking, risk decision making, sensorimotor ability, and vigilant attention.
- Visual Analog Scales: An 11-point sliding scales that measure psychosocial constructs such as mood, fatigue, conflict, and stress.
- Behavioral health and human factors questionnaire: This questionnaire includes both "post-sleep" and "pre-sleep" questionnaires with items focused on sleep (e.g. amount, quantity, quality), mood, affect, team cohesion and performance, and crew living/habitability within the International Space Station (ISS) vehicle. A preflight personality survey is also given.
- Actigraphy and sleep-related questions: Includes data downloaded from an advanced technology actigraphy wrist watch and includes data about lighting. On the ground, a brief post-sleep survey is completed daily during the two-week periods over which actigraphy is collected. This survey includes questions related to sleep quality. In flight, this same survey is filled out along with an additional pre-sleep questionnaire containing questions related to mood, workload, and crew living. Both are to be completed once per month during flight.
- ROBoT: The Robotic On-Board Trainer (ROBoT) is a training simulator that is used by astronauts to rehearse docking and grappling maneuvers using the robotic arm on the space station. The simulation is based on highly realistic 3D simulations of the robotic arm on the ISS and associated physics relating to spaceflight. The ROBoT system simulation involves a difficult and critical spaceflight maneuver of docking an incoming spacecraft. To complete the task, the participant must extend the robotic arm to the incoming spacecraft, line the end effector up with a target on the approaching vessel, and grapple a pin on the vessel to "capture" the target. This maneuver requires situation analysis, planning, decision-making, object orientation, mental rotation, visual processing, fine motor control, and visual motor integration (Johnson & Alexander, 2013). A computerized research adaptation of ROBoT has been developed for use as an operational task to assess human performance. This ROBoT-r tool could help identify the optimum relationship between crew training, practice, and actual operations and identify performance measurements that could make the astronaut interaction with the system more effective and more efficient (Ivkovic et al. 2019; Vos et al., 2017).

3. Human Factors Standard Measures

General human performance measures exist that are well known, reliable, and validated. For example, measures of accuracy and time when performing a task have been widely used as measures of human performance for many decades. However, the challenge is that their implementation and interpretation usually must be tailored for individual tasks, operations, systems, etc.

A survey of NASA and the Department of Defense (DoD) subject matter experts (SMEs) uniformly recommended standard measures related to various forms of accuracy and time and ***note that collection of such data should be built into onboard systems for all future NASA manned missions, including LDEMs***. They also recommend measures that assess workload and situational awareness. For example, Dr. Gordon Vos of NASA JSC (personal communication) states that:

> Time on task, error rates, workload, and situation awareness are basic and yet very useful measures for any given task that crew may perform. They are useful in design, in operations, and in research on human performance…Time on task and error rate are best done within a human computer interface [integrated within onboard systems], automatically logged and calculated, with results available for data transmission or periodic download. Workload, and situation awareness are currently only available as survey methods, and their real-time assessment is an active field of current research…These either exist but their collection is not integrated within most systems currently in use on ISS or, in the case of workload and SA, their collection is not yet possible in real-time or unobtrusively, and funded research in those areas would be useful.

This suggests that a priority of future HERA and other analog missions should be to implement such measures in any technological infrastructure (e.g., Pro-X, see below) utilized in experiments conducted in HERA, or in preliminary ground-based experiments simulating similar conditions.

Both accuracy and time can be operationally defined in a number of ways that depend on the nature and complexity of the task and the timeline over which performance is to be assessed. Further, except perhaps for the extremes of performance such as very high or low error rates, it is not always clear how to define thresholds for whatever performance level is deemed acceptable or unacceptable. In future work, it will be important is to establish baselines and criteria for determining when performance is significantly worse that will necessarily be task-dependent. However, for simple repetitive tasks based largely on sensory motor limitations such as typing or interacting with Playbook (Marquez et al. 2013; 2017), establishing a more generalizable baseline and criteria that apply to similar tasks or applications may be possible.

3.1. Accuracy

Accuracy is a measure that assesses the degree of success or failure when performing a task and assumes that success/failure can be well defined (Gawron, 2008). Accuracy may be divided into measures that either characterize successful or correct performance of tasks or those that measure failure or errors. Accuracy measures that focus on success include task completion, percent correct, number correct, correctness score, average score, and probability of correctness (Gawron, 2008; Cuevas, Velasquez & Dattel, 2018). Measures of accuracy that focus on failures or errors are more numerous since there are more ways to fail than to be correct. Examples of such measures include error rate, error number, percent error, absolute error, relative error, deviations, root-mean-square error, false alarm rate (detecting a signal when one is not present), miss rate (not detecting a signal when one is present), and probability of error (Gawron, 2008; Cuevas et

al., 2018). For a complex task, e.g., a procedure execution, measures may include both shorter time scales (accuracy when completing each step of a procedure) and longer time scales (successful completion of the entire procedure).

Measures of an individual's efficacy examine both successes and failures and may use combined metrics such as the ratio of the number of errors divided by the number of correct responses (or vice versa). For example, in a typing training task, a metric formed by the number of errors divided by the number of correct responses demonstrated significant differences between training methodologies and order effects (Ash and Holding, 1990; see Gawron, 2008, p. 21). Since typing is a task in which mistakes can be corrected and overall success depends on both the correct typing of letters and the correction of mistakes, such a ratio metric is more appropriate than measuring either mistakes or successes alone.

Advantages of accuracy measures include the fact that they are objective measures of performance and do not involve speculation about an operator's intention or state of mind. They are also applicable to a wide variety of tasks and independent variables and have been shown to be reliable, valid, and sensitive to independent variables like type of task, display characteristics, environmental stressors, training effects, degree of vigilance, etc.

Accuracy measures may have different statistical strengths or limitations that require different analysis methods. Ratio scale measurements of accuracy (distance from a target) are mathematically robust, enabling the use of parametric analysis techniques that generally assume the underlying response distribution is normal. However, distributions of measures like number of errors, number correct, or percent correct may be skewed (e.g., the presence of floor or ceiling effects), and require mathematical transformation to more closely approximate a normal distribution for application of some types of statistical analyses (e.g. ANOVA).

Additional issues in measuring accuracy may also arise. For example, measuring high accuracy or low error rates can be difficult in practice. An analysis by Ahumada, Beard & Null (2017) found that to be 99% confident that the error rate is less than 1% requires at least 459 observations in a study. If these are low-frequency events, the measurement problem is exacerbated. Also, errors can be of omission (leaving a task out) or commission (doing a task but incorrectly), phenomena that cannot be captured by percent correct. Another issue is the possible presence of speed accuracy trade-offs (SATs), particularly in perceptual-motor tasks, that may occur when an individual trades greater accuracy for a slower response time, or vice versa (see section on Time Measures).

Most importantly, the specific way that accuracy is measured and success or failure is operationally defined should be determined by the specific tasks and research questions involved. Global accuracy measures, independent of individual tasks, operations, systems, etc., probably cannot be achieved.

3.2. Absolute Error and Root Mean Square Error

The mean absolute error (MAE) and the root mean square error (RMSE) are often used as measures of accuracy in tracking tasks such as in control of a robotic arm or remote surface operations. The MAE measures the average magnitude of the errors (e.g., deviation between the correct tracking path and a participant's actual tracked path) without considering their direction (i.e., using absolute values). The RMSE is a quadratic scoring rule that also measures the magnitude of the error. It is the square root of the average of squared differences between the correct track position and the participant's actual tracked position.

Both the MAE and RMSE define average tracking error in units of the variable of interest, e.g. centimeters, with lower values indicating more accurate performance and zero indicates perfect accuracy. The RMSE is particularly appropriate when large deviations are undesirable; it gives a relatively high weight to large errors because the errors are squared before averaging. While the MAE may be the same under various tracking conditions, the RMSE will increase as the variance associated with the frequency distribution of error magnitudes also increases.

The RMSE can be applied and combined across multiple dimensions, e.g., in 3-dimensional (3D) tracking. The RMSE has also been used in assessing such factors as the impact on tracking ability of cueing and environmental effects such as gravitational forces, sleep loss, and types of cockpit displays (see Gawron, 2008, pp. 21–21).

The MAE and RMSE may also be utilized when estimates of average model prediction error are of interest for other variables of interest besides tracking or distance. For example, they can be used to assess how well computational models of performance explain empirically observed behavior. In other words, they measure the quality of the fit between the actual data and the model predictions.

3.3. Time

Time measures are generally based on duration or speed depending on the time frame of interest for the task being measured. Such measures require that tasks have a well-defined beginning and end so that duration can be measured; for example, the total time that it takes for a person or a team to complete a task. Typical duration measures include time on task, task completion time, response time, duration, looking time, movement time, recognition time, and others (Gawron, 2008). For a complex task, such as a procedure execution for inflight maintenance onboard a spacecraft, measures reflecting both shorter (time to complete each step of the procedure) and longer time scales (time to complete specific subtasks like find a stowed part, or the entire procedure) may be of interest.

The terms “response time” and “reaction time” are often used interchangeably but there is a small difference in meaning. “Reaction time” is often associated with the limits of sensory ability, as in making a fast, predetermined response such as a button press to the onset of a visual stimulus as in a Psychomotor Vigilance Task (PVT), while “response time” more generally describes the time to make an overt action in response to a stimulus (Luce, 1986; DeBoeck & Jeon, 2019). Significant increases in time metrics can be sensitive measures of degradations in readiness to perform (e.g. fatigue, attention deficits), but require that well established baselines specific to the task have been determined prior to interpretation.

When efficiency is of interest, measures of time can be combined with accuracy metrics to compute performance speed. A simple example is a typing task in which efficiency is defined as the number of correctly typed letters in a given period of time. Speed and other efficiency metrics are particularly important in domains relevant to space missions such as equipment assembly, maintenance, training, and physical performance. Recently, Vandierendonck (2017) investigated more complex integrated measures and found that two such integrated measures, the linear integration of speed and accuracy (LISA) and the rate correct score (RCS) were the best at detecting independent variable effects and accounting for a larger proportion of the variance in the data.

Dr. Jessica Marquez, Discipline Scientist for the HRP’s Risk of Human and Automation/Robotic Integration (personal communication), notes that there is no measure that “overall quantifies crew

performance—not task specific but overall. I think we can achieve this (carefully as crew does not like to quantify their performance) by understanding how efficiently crew is completing assigned tasks."

Speed accuracy trade-offs (SATs), particularly in perceptual motor tasks, may also occur in which the individual trades greater accuracy for a slower response or vice versa (Heitz, 2014). The nature of such a trade-off for a given task may change depending on factors like instructions (respond as accurately as possible vs. respond as quickly as possible), degree of experience or practice, stimulus quality or discriminability, task difficulty, etc. (Liu & Watanabe, 2012). The SAT has also become a topic of considerable interest in neuroscience. For example, several brain-imaging studies indicate that instructing participants to respond more quickly raises the baseline activity of specific brain regions such as the pre-supplementary motor area and the striatum, with no changes in early sensory or primary motor areas (Bogacz et al., 2010).

3.4. Workload Measures

While human performance measures such as accuracy and time are often used to assess workload, other workload measures incorporate other performance characteristics such as the task load, degree of expended effort, and perceived difficulty (Cuevas et al., 2018). The task load is defined by the total set of goals to be achieved within certain time or resource constraints. Expended effort and the perceived task difficulty are influenced by the nature of the specific tasks to be performed as well as the information and equipment provided by the task environment. Perceived workload may also be affected by individual differences in performers' background knowledge and experience, the strategies adopted to complete the task, and an operator's emotional or cognitive style (Gawron, 2008).

Under high workload, operators tend to hurry when performing a task, resulting in increased errors and poor accuracy. Operators may also exhibit frustration, fatigue, and poor situation awareness of their surroundings. Interestingly, very low workload often produces similar behavior, resulting in high error rates, frustration, fatigue, and poor situation awareness due to boredom, inattention, and complacency from too little to do (Casner & Gore, 2010).

According to Cuevas et al. (2018), workload measures generally fall into four categories:

- Stand-alone (primary task) direct measures
- Secondary task indirect measures
- Subjective measures
- Physiological measures

3.4.1. Stand-Alone (Primary Task) Direct Measures

Standalone or direct measures evaluate workload via performance (e.g., accuracy, time) of the specific task that is of interest. Such an approach assumes that an operator's performance is likely to degrade as workload increases. If performance is acceptable, workload is then assumed to be acceptable. A primary advantage of this method is its simplicity in that one need only observe operator performance. A disadvantage is that speed and accuracy may be insensitive to the state of the operator who may feel overloaded even though performance measures are acceptable. Further, O'Donnell and Eggemeier (1986) have described four potential problems with using task performance as a direct measure of workload. These include the fact that low workload may enhance performance; high workload may result in a floor effect beyond which performance cannot be degraded; there may be confounding effects of training, experience, and

information processing strategy; such measures are necessarily specific to the task an cannot be generalized to other types of tasks.

Another direct measure of workload measures activity by observing the steps and actions that the operator takes in performing a task. The underlying assumption is that a large number of steps or actions implies high workload while few steps implies low workload. The types of steps or actions may include control inputs, verbal responses, mental calculations, decisions, and gazes or visual searches. Again, an advantage of this approach is simplicity. Disadvantages include that a task that requires a small or large number of steps to complete does not necessarily mean that the operator will experience a feeling of being underworked or overworked. The method also ignores the roles of the difficulty of steps and operator skill differences. One example of such an activity measure is the Aircrew Workload Assessment System (AWAS) (Davies et al. 1995; Reid & Nygren, 1988; see Gawron, 2008) developed by British Aerospace that uses time-line analysis software to predict workload and associated error rates for an airplane crew while flying its aircraft. Also, see Gawron (2008, pp.88–92) for descriptions of other types of direct measures of workload.

3.4.2. Secondary Task, Indirect Measures

One of the most widely used methods to assess workload is an indirect or dual-task technique that introduces a secondary (often unrelated) task to the primary operational task environment. Decrements in performance in the secondary task are then taken as an indirect measure of increased workload. Many different types of secondary task measures have been utilized in the literature, such as simple reaction time to a visual or auditory stimulus, the Sternberg memory task, or mental mathematics (Gawron, 2008).

One advantage of the secondary task technique is that it can provide a sensitive measure of operator performance and the ability to distinguish between different equipment/task configurations that would be impossible with only a single task. It can also provide a sensitive measure of task degradation due to stress and provide a common metric for comparing different primary tasks. A disadvantage of the technique is that it relies on assumptions about how secondary task performance competes for the resources (sensory, cognitive, attentional) required for the primary task performance. Other potential problems include: an operator may perform well on the secondary task while performance of the primary task degrades; if using the same secondary task to compare two different primary tasks, it may be unclear how the secondary task overlaps with any given primary task; and operators can have different skill levels, or use different strategies to perform either the primary task, the secondary task, or the combination of the two tasks.

3.4.3. Subjective Measures

Subjective measures of workload ask the human operator to describe the workload they experience when performing a task. They do not attempt to measure anything about the task being performed or the user's performance and depend entirely on the human operator's feelings about their workload. A general limitation of subjective measures of workload is that like all rating scales they require non-parametric analysis techniques. The simplest and least intrusive subjective numerical workload measure is the Instantaneous Self-Assessment (ISA) technique in which subjects rate their overall workload, at periodic intervals, on a scale from 0 to 100. The main advantage of ISA is that it is among the simplest measures to collect. Principle disadvantages arise from differences in the way people think about workload, e.g., physical effort vs. stressful mental effort vs. time-pressure, as well as scale-loading, i.e., differences in the ways individuals interpret and use the 100-point scale (subjects may not use the full scale range to assign ratings, a rating of 50 may not mean mid-level workload for different subjects, etc.).

3.4.3.1. NASA-TLX

The NASA Task Load Index (NASA-TLX) (Hart, 2006; Hart & Staveland, 1988) is one of the most widely used subjective method and has been shown to have high reliability and has been validated under a wide variety of conditions (Gawron, 2008). It was designed to mitigate the disadvantages of scales like ISA. Rather than asking participants to subjectively rate their workload using a single scale, participants must subjectively rate their workload along six different workload sub-scales each designed to characterize workload in a different way: Mental Demand; Physical Demand; Temporal Demand; Performance; Frustration; and Effort. The individual ratings are then combined to form a seventh measure of overall workload weighted by participants' rankings of the degree to which each of the six sub-scales better characterizes their concept of workload. Thus, an advantage of TLX is that it accommodates different ways of conceptualizing the notion of workload.

TLX offers the flexibility of collecting workload measures while participants perform the task or immediately after completion of a task when the operator's memory of the task experience is still fresh. TLX workload ratings can be collected from participants verbally, using a pen and paper, or by computer interface. For example, NASA Ames Research Center (ARC) has recently developed a NASA TLX application that can be downloaded from the Apple store for any iOS device such as an iPhone or iPad. TLX also attempts to mitigate biases about workload that might be due to operators' perceptions of their own performance, e.g. poor performance is interpreted as high workload. Among the disadvantages of the TLX method is that the paper and pencil version is more time-consuming than other rating techniques (TLX for iOS greatly improves administration, scoring, and interpretation time). It also suffers from the same "scale loading" problems as ISA does: operators do not always think of the middle of the rating scale as the medium workload and move linearly toward the two ends of the scale as perceived workload rises and falls.

3.4.3.2. Bedford Workload Scale

The Bedford Workload Scale also collects subjective ratings of workload from participants and is often favored by researchers in the aviation and space communities. It is a 10-level rating scale and offers some of the simplicity of the ISA. The Bedford scale attempts to mitigate the scale-loading problems associated with the ISA and TLX techniques by attaching detailed verbal descriptions to each of the 10 values along the scale. To simplify the process of choosing one of the ten workload ratings, the Bedford juxtaposes a hierarchical decision tree onto the ten scale ratings. Participants navigate through the hierarchy, narrow down their choices of workload ratings to two or three choices, and then select a final single rating based on the descriptions attached to the ratings. Important advantages of the Bedford technique are that it associates descriptions with each of the values along the 1 to 10 scale and the descriptions themselves represent interpretations of the ratings offered by operators. That is, if an operator offers a rating of 7 on the Bedford scale, the text description that is associated with that rating provides its own interpretation of the rating (Casner & Gore, 2010).

The Bedford scale is considered reliable but has not always been shown to be a sensitive measure under conditions such as changes in cockpit display formats, differences in flight control configurations, or differences in combat countermeasure conditions. Also, a survey of Air Force pilots regarding the terminology used in the Bedford scale indicated confusion about the meaning of the different workload ratings (see Gawron, 2008). Another limitation of the Bedford scale is that as operators become proficient with the scale, they report they no longer use the hierarchical choices and proceed directly to the ten ratings; that is, they "memorize" the scale categories without performing the rating technique as designed. Further, the Bedford scale asks subjects to make

judgments about the notion of "spare capacity." Similar to the ambiguities introduced by presenting the word "workload" to subjects, the phrase "spare capacity" can be interpreted as a situation in which the operator variously has additional time, additional mental capacity, or a free hand, etc. An operator that thinks of spare capacity as one of these concepts might give very different ratings than one who thinks of spare capacity differently (Casner & Gore, 2010).

3.4.3.3. Modified Cooper-Harper Scale

The Modified Cooper-Harper Scale also collects subjective ratings of workload from participants and is based on the Cooper-Harper scale originally developed to evaluate aircraft handling qualities. Like the Bedford scale, it uses a 10-point scale superimposed on a decision tree structure. It is designed to assess workload associated with cognitive functions such as perception, monitoring, evaluation, communications, and problem-solving. (Gawron, 2008, p. 167.) Compared to the original Cooper-Harper Scale, scale modifications include asking pilots to rate mental workload level rather than aircraft controllability and emphasizing difficulty or effort rather than control deficiencies. Rating-scale endpoints were also changed to from "excellent, highly desirable" vs. "major deficiencies" in aircraft characteristics to "very easy" vs. "impossible" in terms of required mental effort to complete a task. The modified scale also defined minimal mental effort and adequate performance at lower rating levels than the Cooper-Harper.

Like the Bedford Scale, an advantage of the Modified Cooper-Harper Scale is that it attempts to mitigate scale-loading by providing detailed descriptions for each of the 10 possible ratings. It is also considered a sensitive, valid and statistically reliable measure of overall workload under a variety of types of workload (e.g., communications, navigation, or poor weather conditions) and task conditions (e.g., different cockpit controls/displays). Similar to the Bedford scale, a disadvantage of the Modified Cooper-Harper Scale is that it produces a single rating number that does not measure different aspects or types of workload. Also, as operators learn the scale they "memorize" the scale categories without performing the hierarchical rating technique as designed. Research also suggests that the Modified Cooper-Harper Scale is not as sensitive or as operator-accepted as the NASA-TLX (see Gawron, 2008, p. 169).

3.4.4. Physiological Measures

The fourth group of workload measures is physiological. Some examples include heart rate and heart rate variability, blood pressure, evoked potentials, electrodermal activity, and brain imaging techniques using functional Magnetic Resonance Imaging (fMRI) or functional Near-Infrared Spectroscopy (fNIRS). Physiological measures of workload attempt to associate physiological changes with levels of workload (e.g., see Aghajani, Garbey, & Omurtag, 2017). The goal is to find physiological measures that represent an objective workload measurement without relying on assumptions about how people perceive workload or on their subjective ratings. While a primary advantage of physiological workload measures is their potential to be measured unobtrusively, a primary disadvantage is that there is little underlying theory of physiological mechanism that supports their relationship to workload. Although many physiological measures have been investigated, none has yet proven to definitively demonstrate physiological signatures for the experience of workload. Researchers have observed associated changes in the cardiac, respiratory, and central nervous systems while operators work and have made sensible hypotheses about why these changes might occur. However, there is no clear-cut mechanism by which the same physiological changes should occur in every operator as they perform work (Casner & Gore, 2010).

3.5. Situation Awareness

Situation awareness (SA) has been defined by Endsley (1995) as "the perception of the elements in the environment within a volume of time and space, the comprehension of their meaning and the projection of their status in the near future." Thus, Endsley (1995) specifies three levels of situation awareness: perception of the elements in the environment within a volume of space and time, comprehension of their meaning and current status, and projection of their future status.

The underlying assumption is that greater knowledge along each of these components generally results in improved performance. In general, measures of SA investigate the operator's knowledge about the task being performed, including awareness of an ongoing operation or activity, the environmental context, the states of the various human and/or automated agents involved, and expectations about future state. Objective measures of SA such as probe techniques can be used to directly measure knowledge in real time in real operations as well as in simulations (Cuevas et al., 2018). Subjective measures of SA using rating scales have also been developed.

3.5.1. Situational Awareness Global Assessment Technique

The most frequently used direct measure of SA is the Situational Awareness Global Assessment Technique (SAGAT), which was designed based on real-time, human-in-the-loop simulation of military cockpits but can be applied to other complex human-machine systems. SAGAT is a freeze probe-based technique that involves interrupting performers during a given task and then asking them probing questions about events, objects, cockpit display readings, etc. at that point in time (Endsley, 1988). Subjects' answers are compared with the correct answers that have been simultaneously collected in a computer database during the course of the simulation/task. According to Endsley (1988) the comparison of the real and perceived situation provides an objective measure of SA that can be computed and analyzed in terms of errors and percent correct. SAGAT has been shown to be valid and reliable and can be used in real-world and simulated scenarios (Gawron, 2008). However, Sarter and Woods (1991) have suggested that it does not actually measure SA but rather measures pilot's recall. Others have identified two major problems with objective measures of SA like SAGAT: (1) decay of information (memory decay) and (2) subjects' inaccurate beliefs about events (Fracker & Vidulich, 1991). In general, these issues are problematic for complex tasks like flight since measures such as SAGAT are posing questions to conscious thought when much of the relevant processing actually occurs at unconscious or automated levels and may be largely inaccessible to the pilot.

3.5.2. Situation Present Assessment Method

Another example of a real-time probe-based technique is the Situation Present Assessment Method (SPAM) (Durso et al., 1998) developed to assess air traffic controllers' SA. It is based on the assumption that SA involves knowing where to find information in the environment in order to find a particular piece of information, as opposed to using memory alone regarding what that piece of information is. A set of SA queries are administered online during task execution, but without freezing the task. Subject matter experts prepare queries either before or during task execution, and administer them at the relevant points while the participant is performing the task. The answers and response time are recorded to measure the participant's SA score. Real-time probe techniques can be applied 'in-the-field' and reduce the level of intrusion imposed by task freezes in the freeze-probe techniques. However, the extent to which the intrusion is diminished is questionable because the SA queries are still conducted online during task execution, which signifies a level of intrusion upon the primary task. The SA queries may direct participants to relevant SA information, leading to biased results (Bacon & Strybel, 2013). Furthermore, it is difficult to apply these techniques in dynamic and

unpredictable environments because SA queries must be generated in real-time and that potentially imposes a great burden upon the SMEs (Nguyen et al., 2018).

3.5.3. Situational Awareness Rating Technique

A well-known example of a subjective measure of SA is the Situational Awareness Rating Technique (SART) (Taylor, 1990). It is a self-rating questionnaire method that emphasizes measuring the operator's knowledge in three conceptual areas: (1) demands on attentional resources; (2) supply of attentional resources; and (3) understanding of the situation. Specifically, the SART consists of nine, 7-point rating scales (1 = low; 7 = high) with three subtopics related to each of the three concepts. Advantages of SART are that it is easily administered and is sensitive to factors like participants' performance in a variety of tasks, pilot experience, and fight display characteristics. However, some researchers have found that the three scales were not always uniformly useful or consistent (Gawron, 2008). SA ratings may correspond to performance in a selective manner, i.e. subjects performing well in a trial normally rate their SA as good, while subjects are likely to forget the periods they have poor SA, and more readily recall the periods when they have good SA (Endsley, 1995). Moreover, post-trial questionnaires can only measure SA of participants at the end of the task because humans are normally poor in recalling details of past mental events. Self-ratings are sensitive to individual bias, e.g. subjects often rate poor SA inaccurately as they may not know that they suffer from poor SA from the beginning (Nguyen et al., 2018). Further, since data are ordinal, they must be analyzed appropriately using non-parametric statistics.

Psychophysiological measures of situation awareness have also been investigated (e.g., French, Clarke, Pomeroy, Seymour, & Clark, 2007; Koester, 2007; Vidulich, Stratton, Crabtree, & Wilson, 1994). However, as for workload, such measures are indirect and can be influenced by external environmental factors and operators' individual differences.

3.6. Trust in Automation

A primary underlying justification for automation is that it reduces human effort and improves performance when performing a task. However, Dzindolet et al. (2003) and others have demonstrated that increases in automation implementation do not necessarily lead to corresponding gains in task performance. One reason for this disparity is that performance is impacted by many human factors in addition to an operator's trust in the system. The best performance is produced when operators understand how to appropriately trust and appropriately rely on the automation. Lee and Moray (1992) define trust in automation as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability." Muir's (1994) model of human trust further differentiates between three distinct components of trust: confidence, predictability, and accuracy. The operator has a level of confidence in their prediction of what the system will do. Predictability determines whether an operator can make a prediction about future system behavior. Prediction accuracy can be assessed by comparing the prediction with how the system actually behaved. All should be considered when attempting to measure a user's trust in a system (Cuevas et al., 2018).

Trust has been traditionally assessed by measuring compliance with automation or with a single question asking operators to rate their trust on some scale. Recently, more comprehensive assessments like the *Trust in Automated System Scale* (TASS) (Jian, Bisantz, & Drury, 2000) and the *Human Computer Trust Scale* (HCTS) (Madsen & Gregor, 2000) have been developed. Both scales utilize ratings of a number of statements related to trust to get a more comprehensive picture of trust. The TASS consists of 12 negative and positive statements, e.g., "The system is deceptive" vs. "I am

confident in the system." The 12 statements were derived from word elicitation studies using factor and cluster analyses of words related to human trust. In the resulting scale, participants are asked to rate how well a given statement describes their feeling or impression regarding the automation system being evaluated. A 7-point Likert scale is used ranging from 1 defined as "strongly disagree" to 7 defined as "strongly agree." The HCTS contains 25 positive statements ("The system performs reliably") with 5 statements for each of five constructs related to trust: reliability, technical competence, understandability, faith, and personal attachment. Dolgov & Kaltenbach (2017) directly compared the output of the two scales (TASS and HCTS) for participants using an automated coffee maker. Both scales demonstrated internal consistency. The aggregated scores from both of the measures were found to be significantly correlated with each other. However, they were also differentially sensitive to various aspects of trust as represented by the five constructs of the HCTS (Dolgov & Kaltenbach, 2017). A useful review of the trust in automation literature can also be found in French, Duenser & Heathcote (2018).

3.7. Tools for Assessing Sensory Motor Ability and Impairment

Other types of measurement tools that may be of use during LDEMs are those that assess sensory motor abilities and possible impairments. In addition to reliably providing baseline measures of unimpaired sensory motor abilities, these tools can provide insight into crew cognitive states and/or limitations induced by factors such as fatigue, sleep deprivation, hypoxia, hypercapnia (excessive CO_2), and drug-induced effects. Two such tools discussed here, the PVT and the Comprehensive Oculomotor Behavioral Response Assessment (COBRA), have important potential for LDEMs since they are reliable and relatively brief and easy to administer.

3.7.1. Psychomotor Vigilance Test

As utilized in the Cognition Test Battery, the 3-minute PVT records reaction times (RT) to visual stimuli that occur at random inter-stimulus intervals (Basner et al. 2015). Subjects are instructed to monitor a box on the screen and hit the space bar once a millisecond counter appears in the box and starts incrementing. The reaction time will then be displayed for 1 second. Subjects are instructed to be as fast as possible without hitting the spacebar without a stimulus (i.e., false starts or errors of commission). Other researchers (e.g., Flynn-Evans et al., 2018) may utilize special purpose PVT devices such as the AMI PVT-192 Psychomotor Vigilance Task Monitor. It is a hand-held, self-contained system that stores repetitive reaction time measurements. There is a liquid crystal display (LCD) on the unit which displays instructions and programmable analog mood scales, buttons for the test selection, a microprocessor controlling the unit, solid state storage, and multiple subject recording capability. The length of each test and the inter-stimulus intervals are fully programmable (https://www.artisantg.com/Scientific/69923/AMI_PVT_192_Psychomotor_Vigilance_Task_Monitor).

According to Basner et al. (2015):

> The PVT is a sensitive measure of vigilant attention and the effects of acute and chronic sleep deprivation and circadian misalignment, conditions highly prevalent in spaceflight. The PVT has negligible aptitude and learning effects, and is ecologically valid as sustained attention deficits and slow reactions affect many real-world tasks (e.g., operating a vehicle). Differential activation to PVT performance has been shown across sleep-deprivation conditions, displaying increased activation in right fronto-parietal sustained attention regions when performing optimally, and increased default-mode activation after sleep deprivation, thought to be a compensatory mechanism.

The PVT+ is an iOS app that is in final development at NASA Ames Research Center for use in bed rest studies of sleep deprivation and circadian misalignment (Flynn-Evans et al. 2018). The advantage or "Plus" of NASA PVT+ is its ability to act as a complete study data-collection tool. Unlike standalone commercial PVT tests, the NASA PVT+ for iOS App has the ability to present multiple different industry-standard questionnaires (currently 36 different forms) to subjects at timed intervals throughout the day, or even over multiple days, for both field and in situ studies. The interconnected nature of NASA iOS apps means that both NASA PVT+ and NASA TLX can work together (as well as be launched from web-based forms) that provides a wider range of data collection capabilities. Data itself is securely stored locally in the apps data storage, as well as providing the option to synchronize with a cloud-based data collection server (launching early 2020, currently available as one-off releases under the Apple TestFlight system (Kenji Kato, personal communication)). Both NASA TLX for iOS and NASA PVT+ for iOS are the first NASA research related iOS Apps to be distributed freely for download through the Apple App store. This provides the ability for a significantly larger number of users to adopt and validate the technologies.

A purpose-built device is also being developed at Ames that will allow the PVT app to be calibrated for the latency of the specific device hosting the app. Accurate measurement of the RT in the PVT is critical to being able to compare measurements between experimental conditions in a single study and across studies. Since latency may vary for different hardware platforms or even for different models within the same platform, it is critical to know this value so it can be taken into account when computing RT.

3.7.2. Comprehensive Oculomotor Behavioral Response Assessment

The COBRA is a behavioral data acquisition and analysis system for detecting and characterizing neuro-functional impairment, including mild-to-moderate traumatic brain injury (Liston & Stone, 2014). This novel technology provides a screening tool to detect oculomotor signatures of neurological impairment or injury (U.S. Patent No. 9,730,582; awarded August 15, 2017) and to use multi-dimensional performance measures to characterize the pattern of observed deficits to assist in identifying its cause (U.S. Patent No. 10,420,465; awarded September 24, 2019). Eye movements are the most frequent (~3 per second), shortest-latency (~150–250 ms), and biomechanically simplest (1 joint, no inertial complexities) voluntary motor behavior in primates, and provide a model system to assess sensorimotor disturbances arising from trauma, fatigue, intoxication, aging, or disease states. The technology runs an efficient 5–7-minute behavioral tracking protocol, consisting of spatially and temporally randomized step-ramp radial target motion, combined with reliable analysis tools to assess a wide range of sensorimotor/autonomic/cognitive responses. It uses a set of visual motion stimuli to simultaneously probe pursuit initiation, steady-state tracking, catch-up saccades, visual direction and speed processing, pupillary light reflexes, and eccentric gaze holding as a means of evaluating cortical, cerebellar, and brainstem function.

To assess various aspects of dynamic visual function including peripheral attention, peripheral spatial localization, perceptual motion processing, and oculomotor/pupillary responsiveness, NASA developed a simple clinical behavioral test that measures and computes two dozen ocular-based, or “oculometric,” measures (Stone, 2017). This multidimensional set of oculomotor metrics provides valid and reliable measures of dynamic visuomotor performance within brain circuits and is proving to be a promising assessment tool of neuro-functional performance and health. COBRA can be used to screen for impairment by comparing the oculometric measures of an individual to a normal baseline population (e.g., Liston, Wong & Stone, 2017) or from the same individual before and after exposure to an adverse condition or stressor (Stone, Tyson, Cravalho, Feick, and Flynn-Evans, 2019), e.g., blast, sports, or accident-related head impact; elevated/reduced G-forces; reduced

oxygen; elevated carbon dioxide; drug or alcohol consumption (Tyson, Feick, Cravalho, Tran, Flynn-Evans, & Stone, 2018ab). COBRA can also be used to monitor performance as it returns to baseline during recovery from the exposure or in response to a therapeutic intervention (Flynn-Evans, Tyson, Cravalho, Feick, & Stone, 2019).

Initial results from a traumatic brain injury (TBI) study are promising, showing that this novel NASA technology can reliably detect mild residual impairment of brain function in "recovered" TBI patients, even in the absence of obvious clinical symptoms (Liston, Wong & Stone, 2017). A recent study of acute sleep deprivation shows that mild impairment due to sleep loss and circadian misalignment can be detected after only a few hours of disrupted sleep and, furthermore, the pattern of impairment can be distinguished from that due to TBI or alcohol consumption (Stone, Tyson, Cravalho, Feick, and Flynn-Evans, 2019), indicating that the constellation of COBRA metrics provides both sensitivity and specificity in its assessment of neural impairment. Other potential applications include readiness to perform assessment, performance impact evaluation of adverse operational environments (hypoxia, hypercapnia, sleep deprivation, elevated intracranial pressure, radiation exposure, blast/vibration exposure, etc.), neurology and ophthalmology testing, and drug screening.

3.8. Operational Tasks

NASA may wish to explore other operational tasks like ROBoT to assess the optimum relationship between crew training, practice, and actual operations and identify performance measurements that could be used to make astronaut interaction with systems more effective and more efficient. For example, it may be possible to expand the research capabilities of two projects currently in use or in development at NASA Johnson Space Center.

3.8.1. ISS Emergency On Board Training Simulator

The Emergency On Board Training (OBT) simulator currently in use on ISS could be adapted to collect accuracy, time, and other data in executing emergency procedures in response to vehicle system emergencies including fire, rapid depressurization, and toxic atmosphere.

3.8.1.1. Current Emergency OBT Simulator Capabilities

Currently the overall Emergency simulator system assumes substantial ground-based training (Figure 1) in a full-scale mockup of the ISS which may not be practical for an operational task used as a human factors standard measure. Emergency training on ISS is executed using the Emergency OBT simulator, a stand-alone simulator based on an iPad interface that allows for multi-team, emergency response refresher training (Figure 2). Adapting the standalone portion of the Emergency OBT for data collection may be useful as a standard measure, although it would require some development effort to be able to conduct research either on ISS, in analog environments, or future Gateway spacecraft (POC: Scott Segadi, [SC-CK211], scott.segadi-1@nasa.gov).

3.8.1.2. Current Emergency OBT Simulator Limitations

The crew can run the Emergency OBT simulator autonomously on an iPad or SSC (ISS laptop). However, the simulation requires real-time instructor support provided by Mission Control Center-Houston (MCC-H) to debrief the training event. For autonomous missions, specific "pass/fail' criteria would need to be developed and programmed into the simulator to provide feedback or debrief points.

While the Emergency OBT simulator allows for the crew to traverse ISS while interfacing with virtual hardware and software interfaces and displays via the iPad interface, it does not provide a fully-immersive high-fidelity simulation (e.g., there is no smoke in the cabin, the crew does not close real hatches).

Figure 1. ISS emergency scenarios training. Expedition 46 crew members Tim Kopra of NASA (left) and Tim Peake of ESA (right) engage in an emergency scenario training in the SSTMF responding to smoke in the module. (NASA photo SC2015-E-004471.)

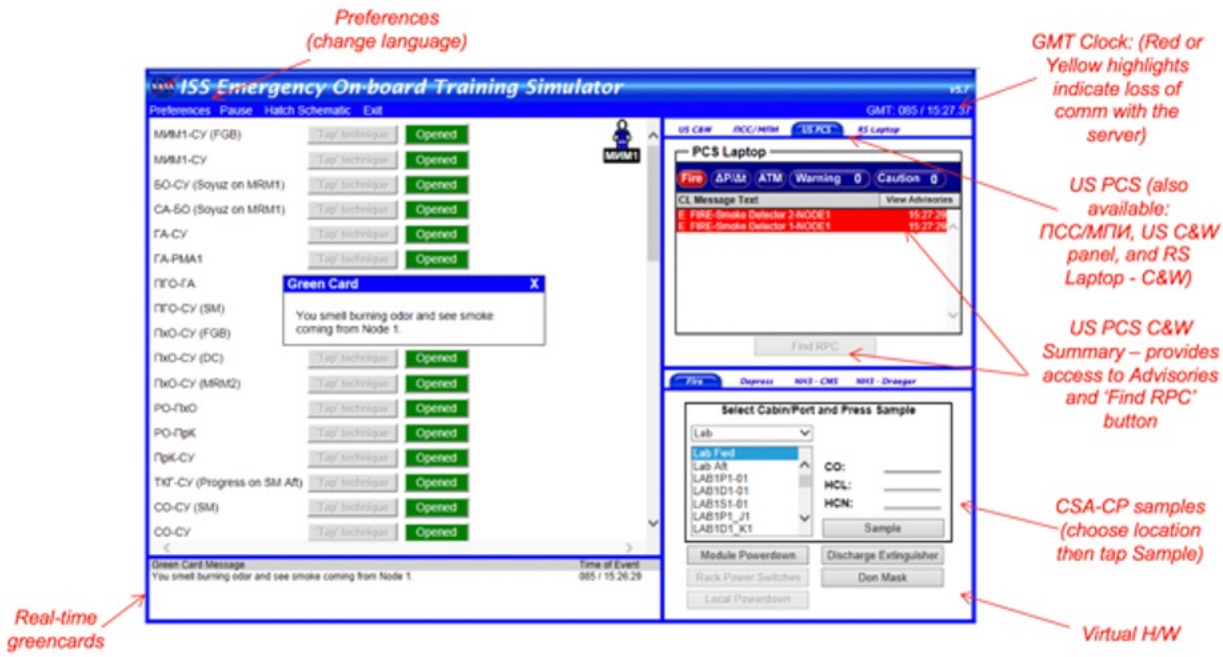

Figure 2. The Emergency OBT simulator. The Emergency OBT simulator provides a crew interface for onboard Emergency Training via an iPad. The interface shown displays hatch status (upper left); messages from the instructor (lower left "green cards"); Caution and Warning messages from the US PCS (upper right); fire port locations (middle right); and icons for virtual hardware (lower right). (ISS OBT Working Group Lead, internal document, personal communication, 2018.)

3.8.1.3. Emergency OBT Simulator Summary

The Emergency OBT simulator provides for ongoing refresher training for emergency response onboard ISS. The simulator is built for the ISS architecture (vehicle layout and volume) and requires ISS Emergency and Warning procedures to execute the training. Adapting the full Emergency simulator for HERA or for a future Gateway vehicle would require onboard resources such as masks, panels, extinguishers, and vehicle-specific procedures. (NASA spent almost 15 years developing integrated emergency procedures for ISS, so developing such procedures for an analog or future vehicle would not be a trivial matter.)

The Emergency OBT simulator is an example of an implementation of ground and on-board training using the latest technology (in this case, an iPad). For use as a human factors standard measure, it may be possible to create a version of the stand-alone OBT simulator software that includes accuracy, time, and other response measurements to provide real-time feedback to the crew as well as record data for later analysis.

3.8.2. Pro-X with Adaptive User Interface Technology and Capability

Future crew will require onboard training for high-risk, critical tasks and complex nominal tasks. Given the expected small size of future vehicles, along with the larger range of tasks that will require in-mission training, NASA will need a platform that provides integration of onboard training needs to meet mission mass and power requirements.

The Pro-X project, formerly named e-Proc, is a NASA Johnson Space Center Engineering Directorate's research project (Wang, 2018; internal NASA document) designed to provide future space flight operations personnel, training personnel, and crew with an electronic procedure platform integrating the entire procedure lifecycle, from procedure authoring and verification to training and execution (Figure 3). Since Pro-X is designed to automatically gather data in real-time on task performance (currently accuracy and time, potentially SA and workload), it has great potential to support research on adaptive, in-mission training and performance support tools. The Pro-X capability for real-time interface adaptation based on human performance measures makes the platform particularly suited as a human factors standard measure for an operational task.

3.8.2.1. Pro-X Capabilities

Pro-X provides a platform to build onboard training across medical, research, and technical tasks; it provides for data gathering of performance measures; it supports adaptation of training and performance support including real-time feedback; and it supports all user interfaces including virtual reality (VR) and augmented reality (AR) devices. Unlike the current ISS procedure platform, Pro-X provides automation and computer oversight during procedure execution allowing for real-time adaption of information to the user based on user-state (e.g., stress, fatigue, workload), the execution environment, or the user preferences.

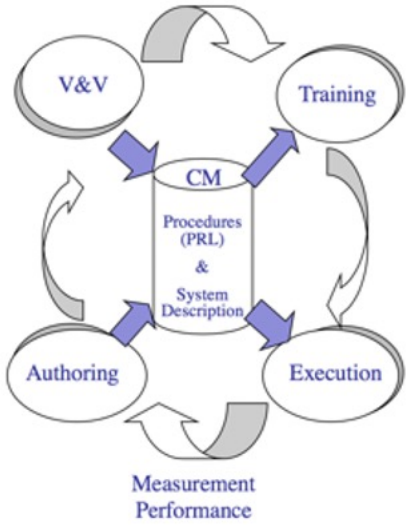

Figure 3. NASA's Pro-X procedure lifecycle development. The Pro-X project will provide a platform to integrate procedure authoring, training, execution, and performance measurement. (Wang, 2018; internal NASA document.)

Pro-X integrates training within the procedure development and execution system. The procedure and training modules developed to date (Figure 4) provide just-in-time training (JITT) during procedure execution. They include medical, research, and technical tasks similar to tasks that future crewmembers will require in-mission onboard training to perform. Prototype training/performance support tools have been developed for the following tasks:

- AR ultrasound
- MiniDNA sequencing
- AR Advanced Resistive Exercise Device (ARED) maintenance
- AR stowage

Pro-X also links to Playbook (Marquez et al. 2013; 2017), allowing astronauts to call up procedures from their timeline schedules as needed.

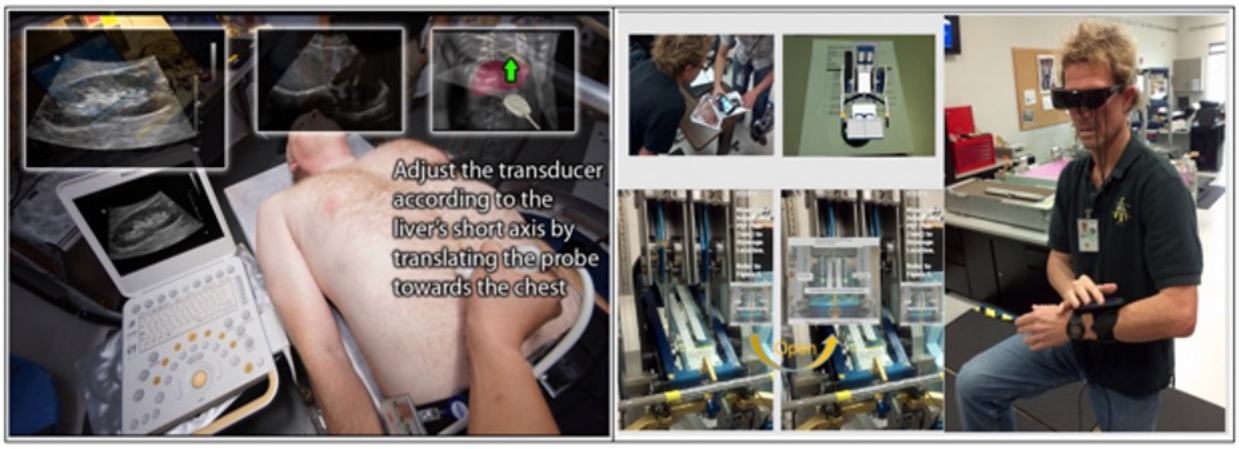

Figure 4. Augmented reality training. AR is used to enhance crew medical training on ultrasound examination (left) and crew training and performance support on ARED maintenance (right). (Wang, 2018; internal NASA document.)

3.8.2.2. Pro-X Limitations

The Pro-X platform is a research platform, not yet deployed in NASA's missions, however, the capabilities provided by Pro-X continue to mature. While Pro-X is designed to adapt to user state, research is needed to determine how best to provide user state information to Pro-X (e.g., stress, fatigue, workload) and to determine how information should then be adapted given the determined state (e.g., reduce information on screen, provide haptic feedback, change font size).

3.8.2.3. Pro-X Summary

Pro-X is a platform for electronic procedures that integrates the entire procedure lifecycle, from procedure authoring and verification to training and execution, designed to allow for adaptive training/performance support. HCAAM researchers working to determine user-state and provide necessary adaptations will require a platform with an easy authoring tool, with the capability to include any type of training content and technology (e.g., video, AR, VR), and with the technology that supports real-time adaptation. The Pro-X platform provides the technology to do so for any task being trained or executed. Providing Pro-X in HERA as well as in future Gateway missions would support researchers investigating adaptive, in-mission initial training, JITT, and performance support tools, all of which will be needed for NASA's future LDEMs. The capability for real-time interface adaptation based on human performance measures makes the platform particularly suited for use as a human factors standard measure. Adapting Pro-X tasks such as miniDNA Sequencing and AR Stowage to the HERA environment should be possible in the relative near-term but will require some time, programming effort, and resources.

3.8.3. Operational Performance Measures: Playbook Data Mining

Jessica Marquez, Steven Hillenius, and John Karasinski at NASA's Ames Research Center have proposed a methodology to identify human performance metrics codifiable from interaction data collected during the crew's frequent use of Playbook, a scheduling and execution tool. Playbook can be used to view and modify daily schedules and assigned activities and select procedures for each

activity. It also allows users to input “crew notes” (written feedback on activities) and status activities (mark them as in progress or completed).

User interaction data automatically collected by Playbook could be used to identify operational performance metrics that are codifiable. Being able to easily codify these metrics from interaction data is relevant for several reasons. First, codifiable metrics can serve as unobtrusive, objective metrics of performance instead of relying on crew reports or surveys. Second, it results in metrics that do not speculate on crew’s intent nor depend on researcher’s subjective evaluation (akin to manual analysis). This is particularly challenging for communication analysis. And third, metrics that can be computed and reliably automatically coded may provide a methodology for longitudinal tracking of real-time changes in human performance. This will lead the way to a non-manual process, which is necessary for high frequency metrics in a longitudinal analysis. The benefits of codifiable metrics suggest that daily trends of human performance could be obtained in the future and be integrated into spaceflight operations to help monitor changes.

3.9. Habitability

The physical environments of space vehicles and habitats can critically affect the health and well-being of crew, and thereby impact mission success. Sources of discomfort due to poorly designed habitats may include inadequate volume in which to live and work, auditory interference with privacy and tasks, olfactory distress, frustration over confusing hardware and software interfaces, and other stressors (Beaubien & Baker, 2002). These issues will be exacerbated by the level of isolation that crewmembers of long-duration spaceflight missions will experience and have the potential to contribute to reductions in crew safety, introduction of inefficiencies and errors, and reduced satisfaction (Greene, Thaxton, & Adolf, 2019).

Documenting and quantifying details about crew task performance and well-being in a long-duration microgravity environment can provide valuable data for use in research, operations and the design of future vehicles. Historically, the ability to capture, analyze and make habitability recommendations based on crew experience and observations has been limited to information collected during post-flight debrief sessions that occur on average within a month of a crew’s return from ISS. Limited additional data is retrieved from information entered into the crew notes system, which is a part of the crew scheduling tool known as Operations Planning Timeline Integration System (OPTIMIS).

More recently, tools intended for the collection of real-time or near real-time human factors data have been developed. Greene, Thaxton, and Adolf (2019) define real-time data collection as observations documented by crewmembers as soon as reasonably possible after an observation is made rather than during post-mission debriefs. A crewmember is not asked to stop a task to document an observation, but the observation is recorded during the mission, ideally while the observation is still fresh on the crewmember’s mind. To this end, the Space Habitability Observation Reporting Tool (SHORT), and later an iOS-based SHORT (iSHORT), were developed to take advantage of advances in technology for enhanced reports (Thaxton, Litaker, Jr., & Toy, 2012). iSHORT provides a simple iPad interface for users to report positive or negative observations about their environment, equipment, and general activities within the habitat. iSHORT has multimedia reporting capabilities, and methods for the collection of video data; the application allows users to report observations related to human factors and habitability using text, photographs, videos, and/or audio recordings.

iSHORT was tested as part of the NASA Extreme Environments Mission Operations (NEEMO) 16 mission, enabling an assessment of how the tools work in an operational environment (Thaxton,

Litaker, Jr., & Toy, 2012). iSHORT was also recently used to assess habitability and human factors on the ISS to better prepare for future long-duration space flights (Greene, Thaxton, & Adolf, 2019). Data collection sessions primarily required the use of an upgraded iSHORT (iPad) application to capture near real-time habitability feedback and analyze vehicle layout and space usage. In addition, a stand-alone Question and Answer application (iQ&A) was utilized with capabilities that were previously part of iSHORT. iQ&A was developed to make questionnaire creation more accessible and to simplify the user experience for iSHORT. It provided a platform for survey administration that took advantage of the multimedia functions available on an iPad. Survey administrators could create questionnaires with a variety of input methods (free response, radio buttons, checkboxes, etc.) and survey takers could complete the questionnaire on an iPad and attach video, photographs, text, or audio files. Both tools are planned to be released for free on the Apple Applications Store.

During ISS data collection, a total of six subjects from five recent ISS missions ranging from standard (~6 months) ISS mission length to 1-year were tested. Participants were asked to capture observations about their environment about once every two weeks; to capture a walk-through video of an area of ISS about once per month; to narrate a task about once per month; to complete a human factors and habitability questionnaire three times per mission; and to participate in a conference with the investigator team following each questionnaire. Content analysis (see Stemler, 2001; Stuster, 2010, 2016) was used to categorize and code the data, draw general conclusions, and make recommendations for future vehicle and habitat design.

The study results demonstrated that habitability and human factors concerns during long-duration microgravity exposure were successfully characterized by the use of the iSHORT and iQ&A tools. The relationships between environmental factors, mission phase, and performance were explored. In general, crewmembers provided thoroughly detailed and insightful feedback for all data collection types. They had a good understanding of the types of details that are of interest to habitability and human factors experts, and provided a wealth of information relevant for the design of future space vehicles and habitats. Providing an opportunity for participants to capture observations while on-orbit allowed them to give demonstrations and discuss details that are fresh on their mind. Ideally, this type of near real-time reporting should continue as part of regular operations on ISS and in future programs, potentially as an integrated part of the Crew Notes functionality. Such data will be of great value to the greater operational community as well as human factors researchers and vehicle designers.

4. Recommendations and Future Work

In light of information presented in this report based on the results of the literature review and interviews conducted with NASA and military experts, general recommendations regarding human factors metrics for use in spaceflight, HCAAM, analog facilities, and other Human Research Program studies include the following.

4.1. General Recommendations

While NASA experts have tremendous knowledge of their current vehicle systems and disciplines as well as the operational concepts that inform current mission needs, they are not as knowledgeable on operational concepts for LDEMs. Further, while the Mars task list (Stuster et al., 2019) provides a detailed listing of tasks for such a mission, the listing does not specifically categorize tasks by risk to mission success and does not indicate the underlying skills necessary to perform these tasks. In order to provide operational experts, HCAAM and other researchers, and the Human Research Program

context for future HFBP research needs and to inform the development of appropriate human factors standard measures it is recommended that:

- HFBP develop a concept of operations (Con Ops) that documents mission assumptions for future Long Duration Exploration missions.
- HFBP develop a list of high-risk, critical crew tasks for future Gateway and DST missions, accepted by the operational community, to be included in the Con Ops.
- HFBP determine the skills necessary to perform these tasks.

Since selection and interpretation of appropriate human performance measures depends on contextual factors such as the specific task, operation, or human-machine system involved, it is recommended that:

- Measurement capabilities for human performance metrics should be developed as an inherent part of the technological infrastructure of analog environments such as HERA and in all onboard systems in future NASA manned spacecraft. One example of this may be the use of a Pro-X based infrastructure for supporting procedure execution, measuring performance, and validating the usefulness of particular adaptive, in-mission training and performance support tools that have been developed.
- It will be important to establish baseline performance levels and normal performance variability in critical tasks as well as criteria for determining when performance is significantly worse or better; these criteria will necessarily be skill or task-dependent and may depend on individual differences among crew members.
- To the extent possible, unobtrusive metrics should be utilized. Accuracy and time measures may be made unobtrusive by including their measurement as an inherent part of onboard systems. However, such metrics will need to be tailored for individual tasks and skills.
- Physiological metrics or oculometrics may be possible unobtrusive metrics, although further work is needed to definitively demonstrate consistent physiological signatures for various human behavioral experiences such as workload or situational awareness. The consistency of such metrics is likely to be dependent in complicated ways on a variety of factors such as fatigue, stress and individual differences in experience or proficiency.
- The judicious use of more overt metrics in the form of surveys and crew reports may remain a necessity since it is unlikely that unobtrusive measures may be developed for monitoring some aspects of crew performance, health and safety. These may include survey-based measures of workload (NASA-TLX, Bedford Workload scale), situational awareness (SAGAT, SPAM, SART), trust in automation (TASS, HCTS), and habitability (iSHORT, iQ&A). However, some of these metrics may be more appropriate in studies supporting initial development of spacecraft systems rather than as a routine capability of onboard infrastructure.
- Tools that assess sensory motor abilities and possible impairments may also be of use during LDEMs. In addition to providing baseline measures of unimpaired sensory motor abilities, these tools can provide insight into crew cognitive states and/or limitations induced by factors such as fatigue, sleep deprivation, hypoxia, hypercapnia, and drug-induced effects. For example, the PVT, the Cognition Test Battery, and the COBRA, have important potential for LDEMs since they are reliable and relatively brief and easy to administer.

- NASA may wish to explore operational tasks to assess the optimum relationship between crew training, performance, and actual operations and identify performance measurements that could be used to make astronaut interaction with systems more effective and more efficient. Given sufficient resources, it may be possible to expand the research capabilities of projects currently in use or in development at NASA Johnson Space Center, e.g., ROBoT-r, the Emergency OBT Simulator, and the Pro-X Adaptive User Interface infrastructure for procedure execution for a variety of tasks.

4.2 Specific Recommendations for HCAAM-VNSCOR and LDEM-Related Studies

- Selection of human factors measures for studies should be based on whether they occur in operational or research and development contexts. Studies conducted in operational settings are likely to be restricted in terms of the allowable degree of obtrusiveness and the time that may be allotted to data collection. Research being conducted to develop and evaluate new concepts and technologies will allow the use of more extensive data collection techniques (see Table 1).
- Use common measures across studies, particularly for HCAAM-VNSCOR (Virtual NASA Specialized Center of Research) studies in HERA.
- Utilize the Principle of Converging Operations[2] when designing studies. For example, whenever possible, collect objective performance measures (accuracy and time) as well as subjective or self-report measures such as workload, SA, and Trust in Automation. This can be more readily and unobtrusively achieved when measurement capabilities for human performance metrics are developed as an inherent part of the technological and research infrastructure of analog environments such as HERA and in all onboard systems in LDEM spacecraft.

4.3. Recommended Measures for Operations vs. Research and Development

Table 1 provides the recommended standard measures useful for operational vs. research and development studies, particularly for HCAAM-VNSCOR studies in HERA. Please see the body of this report for details on individual measures.

[2] When psychologists use multiple operational definitions of the same construct—either within a study or across studies—they are using *converging operations*. The idea is that the various operational definitions are "converging" or coming together on the same construct. When scores based on several different operational definitions are closely related to each other and produce similar patterns of results, this constitutes good evidence that the construct is being measured effectively and that it is useful." (Price et al., 2017, Ch. 4, p. 63).

Table 1: Recommended Measures for Operations vs. Research and Development		
Standard Measure	*Operation Studies, e.g., on ISS*	*Research & Development*
Accuracy*	% correct, other accuracy measures as appropriate to task	% correct, other accuracy measures as appropriate to task
MA error, RMS error*	MAE, RMSE for tracking or similar tasks	MAE, RMSE for tracking or similar tasks
Time*	Response or reaction time, time on task, task/step completion time, other time measures as appropriate to task	Response or reaction time, time on task, task/step completion time, other time measures as appropriate to task
Workload	Accuracy and time used as converging operations for primary, secondary tasks*; use with subjective measures like Bedford Workload Scale that requires less time & effort to administer than NASA-TLX	Accuracy and time used as converging operations for primary, secondary tasks*; use with subjective measures like NASA-TLX, considered more reliable and well-validated
Situation awareness (SA)	SART, subjective measure that requires less time & effort to administer	SAGAT, SPAM: objective measures, both are probe techniques for high fidelity sims that require more time & effort to administer
Trust in automation	Objective measures of automation compliance, time, accuracy* combined with single question rating of trust	Objective measures of automation compliance, time, accuracy* combined with Trust in Automated System Scale (TASS) or Human Computer Trust Scale (HCTS)
Assessment tools for sensory-motor ability	PVT+, Cognition Test Battery, COBRA (likely pre- and post-task use)	PVT+, Cognition Test Battery, COBRA (likely pre- and post-task use)
Operational tasks	Tasks such as ROBoT-r and Pro-X adaptive procedure platform still in development	Tasks such as ROBoT-r and Pro-X adaptive procedure platform still in development
Habitability	iSHORT, iQ&A	iSHORT, iQ&A

* Assumes measures are collected as unobtrusively as possible in that they have been integrated into the technological infrastructure.

References

Aghajani, H., Garbey, M., & Omurtag, A. (2017). Measuring mental workload with EEG+fNIRS. *Frontiers in Human Neuroscience, 11*(359). doi: 10.3389/fnhum.2017.00359 .

Ahumada, A.J., Beard, B.L., & Null, C.H. (2017). Accounting for the Speed-Accuracy Trade-Off in Quantifying Human-In-The-Loop Error Probabilities. *Proceedings of AIAA SciTech Forum and Exposition. Grapevine, TX. January 9-13, 2017.* DOI:10.2514/6.2017-1097 .

Ash, D. W., & Holding, D. H. (1990). Backward versus Forward Chaining in the Acquisition of a Keyboard Skill. *Human Factors*, *32*(2), 139-146. https://doi.org/10.1177/001872089003200202 .

Bacon, L. P., & Strybel, T. Z. (2013). Assessment of the validity and intrusiveness of online-probe questions for situation awareness in a simulated air-traffic-management task with student air-traffic controllers. *Safety Science, 56,* 89–95.

Basner, M., Savitt, A., Moore, T. M., Port, A. M., McGuire, S., Ecker, A. J., … Gur, R. C. (2015). Current Development and Validation of the Cognition Test Battery for Spaceflight. *Aerospace Medicine and Human Performance. 86*(11), 942–952.

Beaubien, J. M., & Baker, D. P. (2002). A review of selected aviation Human Factors taxonomies, accident/incident reporting systems, and data reporting tools. International *Journal of Applied Aviation Studies, 2*(2), 11–36.

Bogacz, R., Wagenmakers, E.-J., Forstmann, B. U., & Nieuwenhuis, S. (2010). The neural basis of the speed–accuracy tradeoff. *Trends in Neurosciences, 33*, 10–16.

Casner, S. M., & Gore, B. F. (2010). *Measuring and Evaluating Workload: A Primer.* NASA/TM—2010-216395.

Cuevas, H. M., Velázquez, J., & Dattel, A. R. (2018). *Human Factors in Practice: Concepts and Applications.* Boca Raton: CRC Press, Taylor & Francis Group.

Davies, A. K., Tomoszek, A., Hicks, M. R., & White, J. (1995). AWAS (Aircrew Workload Assessment System): Issues of theory, implementation, and validation. In R. Fuller, N. Johnston, & N. McDonald (Eds.), Human factors in aviation operations. *Proceedings of the 21st Conference of the European Association for Aviation Psychology (EAAP)* Vol 3.

DeBoeck, P. & Jeon, M. (2019). An Overview of Models for Response Times and Processes in Cognitive Tests. *Frontiers in Psychology, 10*(102). doi: 10.3389/fpsyg.2019.00102 .

Dinges, D. F., Basner, M., Stahn, A., Roma, P., Strangman, G., & Stuster, J. (2017). Behavioral core measures (previously SBMT): Overview of data collection in HERA 30-day missions. *Proceedings of the NASA Human Research Program Investigators' Workshop, A New Dawn: Enabling Human Space Exploration.* Abstract 17125. Retrieved from https://three.jsc.nasa.gov/iws/SRIW-Cvent-Program-2017.pdf .

Dinges D., Basner M., Strangman G., Stuster J., Roma P., Mollicone D., … Williams T. (2018). Standardized behavioral measures for detecting behavioral health risks during exploration (behavioral core measures). *Proceedings of the NASA Human Research Program Investigators' Workshop: The Gateway to Mars.* Abstract 18007. Retrieved from https://three.jsc.nasa.gov/iws/FINAL_2018_HRP_IWS_program.pdf .

Dolgov, I., & Kaltenbach, E. K. (2017). Trust in Automation Inventories: An Investigation and Comparison of the Human-Computer Trust and Trust in Automated Systems Scales. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *61*(1), 1271–1275. https://doi.org/10.1177/1541931213601799 .

Durso, F. T., Hackworth, C. A., Truitt, T. R., Crutchfield, J., Nikolic, D., & Manning, C. A. (1998). Situation awareness as a predictor of performance for en route air traffic controllers. *Air Traffic Control Quarterly, 6*(1), 1–20.

Dzindolet, M., Peterson, S., Pomranky, R., Pierce, L., & Beck, H. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies, 58*(6), 697–718.

Endsley, M. R. (1988). Situation awareness global assessment technique (SAGAT). In *Proceedings of the IEEE National Aerospace and Electronics Conference, NAECON 1988*, 789–795. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

Endsley, M. R. (1995). Toward a theory of situation awareness in dynamic systems. *Human Factors, 37*(1), 32– 64.

Flynn-Evans, E., Wong, L., Pradhan, S., Karasinski, J., Hu, C., Strangman, G., Ivkovic, V., Anderson, M., & Stone, L. (March, 2018). *Performance on the robotic on-board (ROBoT-r) simulator during sleep loss and circadian misalignment.* NASA Human Research Program Final Report, NASA Ames Research Center.

Flynn-Evans, Tyson, Cravalho, Feick, & Stone (2019) Low-dose caffeine administration during acute sleep deprivation eliminates visual motion processing impairment, but does not improve saccadic rate. *Sleep Research Society Annual Meeting (SLEEP 2019), June 08, 2019 - June 12, 2019*, San Antonio, TX; United States.

Fracker, M. L., & MA Vidulich, M. A. (1991). Measurement of situation awareness: A brief review. In R. M. Taylor (Ed.) *Situational awareness in dynamic systems*. (IAM Report 708). Farnborough, UK: Royal Air Force Institute of Aviation Medicine.

Freedy, A., DeVisser, E., Weltman, G., & Coeyman, N. (2007, May). Measurement of trust in human-robot collaboration. In *International Symposium on Collaborative Technologies and Systems*. CTS 2007, 106–114. Piscataway, NJ: Institute of Electrical and Electronics Engineers.

French, H. T., Clarke, E., Pomeroy, D., Seymour, M., & Clark C. R. (2007). Psycho-physiological measures of situation awareness. In J. Noyes, M. Cook, & Y. Masakowski (Eds.). *Decision making in complex environments.* Boca Raton, FL: CRC Press, Taylor & Francis Group.

French, B., Duenser, A., Heathcote, A. (2018). *Trust in Automation—A Literature Review.* CSIRO Report EP184082. CSIRO, Australia.

Gawron, V. (2008). *Human performance, workload, and situational awareness measures handbook.* Boca Raton, FL: CRC Press, Taylor & Francis Group.

Greene, M., Thaxton, S., & Adolf, J. (2019). *Habitability assessment of international space station (ISS habitability).* NASA Human Research Program Final Report, NASA Johnson Space Center.

Guide to Human Performance Measurements (2001). ANSI/AIAA G-035A-2000. American National Standards Institute.

Hart, S. G. (2006). NASA-task load index (NASA-TLX): 20 years later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 50*(9), 904–908.

Hart, S. G., & Staveland, L. (1988). Development of the NASA task load index (TLX): Results of empirical and theoretical research. In P. A. Hancock & N. Meshkati (Eds.), *Human Mental Workload* (pp. 239–250). Amsterdam: North Holland.

Heitz, R. P. (2014). The speed-accuracy tradeoff: History, physiology, methodology, and behavior. *Frontiers in Neuroscience, 8*, Article 150. Published online 2014 Jun 11. doi: 10.3389/fnins.2014.00150 .

Hashemi, S., & Hillenius, S. (2013). "@NASA: The user experience of a space station." *SXSW Interactive,* Austin, TX. Speech.

Ivkovic, V., Sommers, B., Cefaratti, D. A., Newman, G., Thomas, D. W., Alexander, D. G., & Strangman, G. E. (2019). Operationally relevant behavior assessment using the Robotic On-Board Trainer for Research (ROBoT-r). *Aerospace Medicine and Human Performance. 90*(9), 1–7.

Jian, J-Y., Bisantz, A. M., & Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *International Journal of Cognitive Ergonomics, 4*(1).

Johnson G., & Alexander G. (2013). *Robotics on-board trainer (ROBoT).* (Report No.: MSC-25005-1). Houston (TX): NASA Johnson Space Center.

Koester, T. (2007). Psycho-physiological measurements of mental activity, stress reactions and situation awareness in the maritime full mission simulator. In J. Noyes, M. Cook, & Y. Masakowski (Eds.), *Decision making in complex environments* (pp. 311–320). Burlington, VT: Ashgate Publishing, Ltd.

Lee, J. D., & Moray N. (1992). Trust, control strategies and allocation of function in human–machine systems. *Ergonomics, 35*(10), 1243–1270.

Lee, J. D., & Moray N. (1994). Trust, self-confidence, and adaptation to automation. *International Journal of Human-Computer Studies, 40*(1), 153–184.

Lee, J. D., & See K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80.

Liston, D. B., & Stone, L. S. (2014). Oculometric assessment of dynamic visual processing. *Journal of Vision, 14*(14):12, 1–17.

Liston, D. B., Wong, L. R., & Stone, L. S. (2017). Oculometric assessment of Sensorimotor Impairment Associated with TBI. *Optometry and Vision Science, 94*(1):12, 51–59.

Liu, C. C., & Watanabe, T. (2012). Accounting for speed–accuracy tradeoff in perceptual learning. *Vision Research 61*(2012), 107–114.

Luce, D. (1986). *Response times: Their role in inferring elementary mental organization.* New York, NY: Oxford University Press.

Madsen, M., & Gregor, S. (2000). Measuring human-computer trust. *11th Australasian Conference on Information Systems* (Vol. 53, pp. 6–8), Brisbane, Australia: Australasian Association for Information Systems.

Marquez, J. J., Pyrzak, G., Hashemi, S., Ahmed, S., McMillin, K., Medwid, J., Chen, D., & Hurtle, E. (2013). Supporting real-time operations and execution through timeline and scheduling aids. *International Conference of Environmental Systems*, ICES, Vail, CO.

Marquez, J. J., Hillenius, S., Deliz, I., Kanefsky, B., Zheng, J., & Reagan, M. (2017). Increasing crew autonomy for long duration exploration missions: Self-scheduling. *2017 Aerospace Conference,* IEEE.

Mount, F. E. (2002). Habitability: An Evaluation. In H. W. Lane, R. L. Sauer, & D. L. Feeback (Eds.), *Isolation: NASA Experiments in Closed-Environment Living* (pp. 87–116). San Diego, CA: American Astronautical Society.

Muir, B. M. (1994). Trust in automation: Part I. Theoretical issues in the study of trust and human intervention in automated systems. *Ergonomics, 37*(11), 1905–1922.

Mukherjee, S., Yadav, R., Yung, I., Zajdel, D. P., & Oken, B. S. (2011). Sensitivity to mental effort and test-retest reliability of heart rate variability measures in healthy seniors. *Clinical Neurophysiology, 122*, 2059-2066.

National Aeronautics and Space Administration. (2014). Section 5.7 Cognitive workload. (2014). In *Human Integration Design Handbook (HIDH), Revision 1* (pp. 201–236). (NASA/SP-2010-3407/REV1).

Nguyen, T., Lim, C. P., Nguyen, N. D., Gordon-Brown, L., & Nahavandi, S. (2018). A review of situation awareness assessment approaches in aviation environments. *IEEE Systems Journal.* doi: 10.1109/JSYST.2019.2918283 .

O'Donnell, R. D., & Eggemeier, F. T. (1986). Workload assessment methodology. In K. R. Boff, L. Kaufman, & J. P. Thomas (Eds.), *Handbook of perception and human performance.* New York, NY: John Wiley and Sons.

Price, C. P., Jhangiani, R. S., Chiang, I. A., Leighton, D. C. & Cuttler, C. (2017). *Research Methods in Psychology.* (3rd Ed.), Ch. 4, p. 63, Pressbooks, https://opentext.wsu.edu/carriecuttler/ .

Reid, G. B. & Nygren, T. E. (1988). The subjective workload assessment technique: A scaling procedure for measuring mental workload. *Advances in Psychology, 52,* pp. 185–218.

Sarter, N. B., & Woods, D. D. (1991). Situation awareness: A critical but ill-defined phenomenon, *The International Journal of Aviation Psychology, 1*(1), 45–57, doi: 10.1207/s15327108ijap0101_4 .

Schreckenghost, D. (2016). *Standard measures for space human factors and habitability workshop.* Draft Internal Report by NSBRI Human Factors and Performance Team.

Stemler, S. (2001). An overview of content analysis. *Practical assessment, research & evaluation, 7*(17), 137–146.

Stone, L. S. (2017). *COBRA 2.0: Comprehensive Oculomotor Behavioral Response Assessment is a Tool for Detecting and Characterizing Neuro-Functional Impairment, including Mild-to-Moderate Traumatic Brain Injury.* Unpublished NASA Information Handout.

Stone, L. S., Tyson, T. L., Cravalho, P. F., Feick, N. H., & Flynn-Evans, E. E. (2019). Distinct pattern of oculomotor impairment associated with acute sleep loss and circadian misalignment. *Journal of Physiology* (London), 597(17), pp. 4643–4660. doi: 10.1113/JP277779 .

Stuster, J. (2010). *Behavioral issues associated with long-duration space expeditions: Review and analysis of astronaut journals experiment 01-e104 (Journals): Final report*. NASA/TM-2010-216130. Houston, TX: NASA Johnson Space Center.

Stuster, J. (2016). *Behavioral issues associated with long duration space expeditions: review and analysis of astronaut journals, phase 2: Final report.* Houston, TX: NASA Johnson Space Center.

Stuster, J., Adolf, J. A., Byrne, V. E., & Greene, M. (2019, In Press). *Tasks and abilities for the human exploration of mars.*

Taylor, R. M. (1990). Situation awareness rating technique (SART): The development of a tool for aircrew systems design. In *AGARD Conference Proceedings No. 478, Situational awareness in aerospace operations* (pp. 3/1–3/17). Neuilly Sur Seine, France: NATO-AGARD.

Thaxton, S., Litaker, Jr., H., & Toy, K. (September 14, 2012). *Human factors and habitability assessment tool*: *NEEMO 16 report.* (Human Research Project Control Deliverable). Houston, TX: NASA Johnson Space Center.

Tyson, Feick, Cravalho, Tran, Flynn-Evans, & Stone (2018a) Impairment of Human Ocular Tracking with Low-Dose Alcohol, 28th Neural Control of Movement Annual Meeting; May 1–4, 2018; Santa Fe, NM; United States.

Tyson, Feick, Cravalho, Tran, Flynn-Evans, & Stone (2018b) Increased Dependence on Saccades for Ocular Tracking with Low-Dose Alcohol, Society for Neuroscience annual meeting, Neuroscience 2018; November 3–7, 2018; San Diego, CA; United States.

Vandierendonck, A. (2017). A comparison of methods to combine speed and accuracy measures of performance: A rejoinder on the binning procedure. *Behavior Research Methods. 49(2)*, pp. 653–673. https://doi.org/10.3758/s13428-016-0721-5 .

Vidulich, M. A., Stratton, M., Crabtree, M., & Wilson, G. (1994). Performance-based and physiological measures of situational awareness. *Aviation, Space, and Environmental Medicine, 65*(5), A7–A12.

Vos, G., Cross, E. V., & Russi-Vigoya, N. (2017). *Mission process and task risk operational experience (MP task report).* Houston, TX: NASA Johnson Space Center.

Wang, L. (2018). *NASA Electronic Procedure Technology.* Internal presentation to NASA Johnson Space Center, Houston, TX. Unpublished.

Wilson, G. F. (2000). Strategies for psychophysiological assessment of situation awareness. In *Situation Awareness Analysis and Measurement.* Boca Raton, FL, USA: CRC Press, 2000, pp. 175–188.