Characterizing a Biosignal Sensor Suite as an Approach for Astronaut Performance Model Validation

Christine Fanchiang[a]
The Space Research Company, Centennial, CO, 80122, USA

Kimia Seyedmadani[b]
University of Colorado Boulder, Boulder, CO, 80309, USA

Mark Shelhamer[c]
Johns Hopkins University, Baltimore, MD, 21205, USA

Michael Zero[d] and David M. Klaus[e]
University of Colorado Boulder, Boulder, CO, 80309, USA

Ensuring astronauts are adequately equipped and capable of performing necessary tasks is paramount to achieving mission success especially with the increasing levels of autonomy required for long duration spaceflights. While this may be clear and obvious, the necessary path to achieving such a goal is not so apparent. The spaceflight environment introduces a wide array of stressors that can negatively impact human capabilities and performance; consequently, modeling these interactions quickly becomes overwhelming due to the complexity and adaptivity of the human. While various models have been created to predict astronaut performance during spaceflight, collecting the necessary input data to validate the resultant output of these models remains a challenge. Fortunately, ongoing advances in widely available and unobtrusive wearable biosignal sensors may prove valuable in this endeavor. These sensors enable the opportunity to non-invasively measure and monitor a subset of crew physiological status with respect to performance proficiency parameters. To examine the potential applicability of such sensor systems, an ongoing research project is being conducted under the title "*Using a Human Capabilities Framework to Quantify Crew Task Performance in Human-Robotic Systems*". The project is aimed at validating a crew performance framework using methods of non-invasive monitoring of the crewmember. Within the context of this multi-year project, the present work aims to characterize the process used for selecting and evaluating a suite of Non-invasive Biosignal Sensors (NiBS) while performing spaceflight-like tasks that could provide reliable indicators of changes relevant to performance. Specifically, this work presents preliminary engineering efforts to downselect sensors and develop anticipated test and analysis approaches. This effort is funded through the Virtual NASA Specialized Center of Research (VNSCOR) Human Capabilities Assessments of Autonomous Missions (HCAAM) project.

I. Introduction

Monitoring and predicting astronaut performance can help improve the design of a human spacecraft before it launches and during operations if the modeled system can accommodate and react to the information.

But monitoring and predicting crew performance is currently limited by several factors:

[a] Researcher, The Space Research Company, Centennial, CO, 80122. christine@tsrco.com.
[b] PhD Student, Aerospace Engineering Sciences, Boulder, CO, 80309. kimia.seyedmadani@colorado.edu.
[c] Professor, Otolaryngology Department, Baltimore, MD 21205. mshelhamer@jhu.edu.
[d] PhD Student, Aerospace Engineering Sciences, Boulder, CO, 80309. michael.zero@colorado.edu.
[e] Professor, Aerospace Engineering Sciences, Boulder, CO, 80309. klaus@colorado.edu.

1) There are limited existing published models (or frameworks) for including measured values of the crewmember, environment, and operations into a predictive output for which the system can react[1]. Having such a framework validated would help to identify what type of data should be collected during a mission to achieve the desired predictive output.

2) The existing published models for measuring crew performance impacts due to spaceflight, such as NASA's Integrated Medical Model[2], have focused primarily on probabilistic assessment of medical outcomes and consequences of the astronaut during a mission; these models are not yet directed towards understanding how the design of the system (habitat, robotics, or task itself) could be driving the performance changes.

3) While there are numerous works proving the use of psychophysiological measures for real-time assessments of cognitive or psychological load of different tasks on operators[3-6], to our knowledge none have looked at integrating the physiological measures as an indicator for changes to crewmember capabilities or health and status.

4) While there is a wide range in the quality and utility of sensors available on the market, no recommended integrated minimum number suite of sensors could be identified that has been linked together with a method to provide a view of the crewmember's overall operational state.

Therefore, the current research is aimed at addressing these limitations in astronaut performance monitoring and prediction, with a focus on understanding how the task design affects the astronaut's operational state. This objective will be met through a series of experiments between 2021-2023. This paper presents an initial study that was conducted to investigate the feasibility of a selected NiBS suite as a means of validating a modified crew-task performance model. The goal of this current effort is to document and present the systematic process and rationale employed to define a series of experimental tests to validate the crew performance framework and demonstrate how the data will be processed for inclusion into the model.

To address this objective, the paper is split into three sections. First, we discuss the relevant literature that drove the development and modification of a previously designed crew-task performance model used as the backbone for defining the research methodology and setup. Second, the rationale and methodology are described for selecting and characterizing the experimental setup, which included identifying an appropriate NiBS suite and spaceflight-like tasks. And third, in building the discussed rationale, engineering data is presented to highlight findings regarding the test setup that must be addressed before being able to integrate the data into the crew performance model.

II. Crew Performance Model Background

A. Modifications to a crew performance framework

The framework chosen for this project is derived from prior work by Fanchiang[7], describing the relationship between the crewmember's capabilities, the spacecraft environment, and the task operations. The link between the three components relies on the basic element of a crewmember's capabilities, which is constantly changing due to the surrounding environment, tasks performed previously, and the crewmember's health and status. This framework highlights how the measure of a crewmember's capabilities can be used as the basic comparative metric between the spacecraft environment, tasks, and the crewmember's health and status. Using this framework in this ideal way, we would measure as many variables as possible and track the impacts of every change to the crewmember's performance

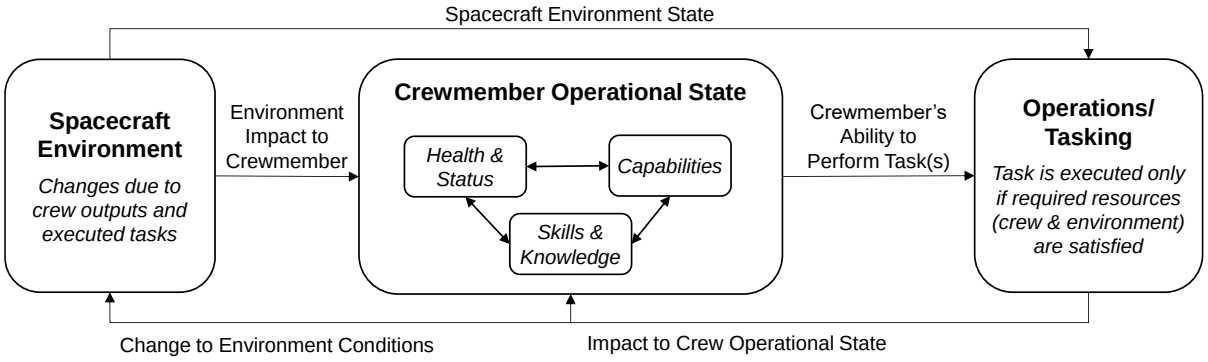


Figure 1. Modified crew performance framework that was originally described in Fanchiang[7].

on various tasks. But in compiling the list of crew capabilities, it quickly becomes obvious that the ability to conduct all the non-invasive data collection is not practical.

While it is possible to measure every crewmember capability listed in previous work from Fanchiang[8], many of the measurements would be highly invasive and take several hours, both of which would affect a crewmember's task execution and generally be highly burdensome. Instead, we expanded the measures to capture this idea of a crewmember's general 'operational' state, which includes 'capability' along with other physiological measures (defined below), which can be measured non-invasively using currently available off-the-shelf sensors.

Crewmember operational state is composed of three elements: the crewmember's health and status, capabilities, and skills and knowledge. Health and status describe the physiological measures of the crewmember such as anthropometrics, current state of the body such as heart rate, core body temperature, etc. Capabilities describe the functions that the crewmember can perform. These functions have modifiers which describe to what degree or level the functions can be performed. The type of capabilities that can be measured and monitored are listed in Fanchiang[8]. These functions vary from moment to moment and can also be modified by the environmental conditions such as cold temperatures reducing fine motor control functionality. Skills and knowledge describe what type of understanding and ability the crewmember has for specific types of tasks. For example, reading is a skill, and it also has modifiers in terms of quality or level. The relationships between capabilities and skills and knowledge are components of learning; capabilities are considered an inherent ability of an able human body, and skills and knowledge have to be learned and can atrophy without practice.

Using this definition for operational state, we can compare the crewmember's operational state before and after a task. This difference is what we define as the expected operational state change resulting from performing a particular task. If the subject then does that same task under a different set of circumstances, we would expect the operational state change to be different, thus informing us of a change in the crewmember's capabilities. For instance, a change in the environmental conditions such as poor lighting or a change in the crewmember's capability such as an injured arm would likely yield unique operational states compared to performance monitored under ideal environmental conditions.

III. Rationale and Systematic Process for Experiment Design

A. Task Selection Process

To set up the experiments to validate this framework, we first needed to characterize and identify the type of tasks we would have subjects perform that both mimicked expected spaceflight tasks and would allow us to use existing sensors for collecting high quality data to analyze.

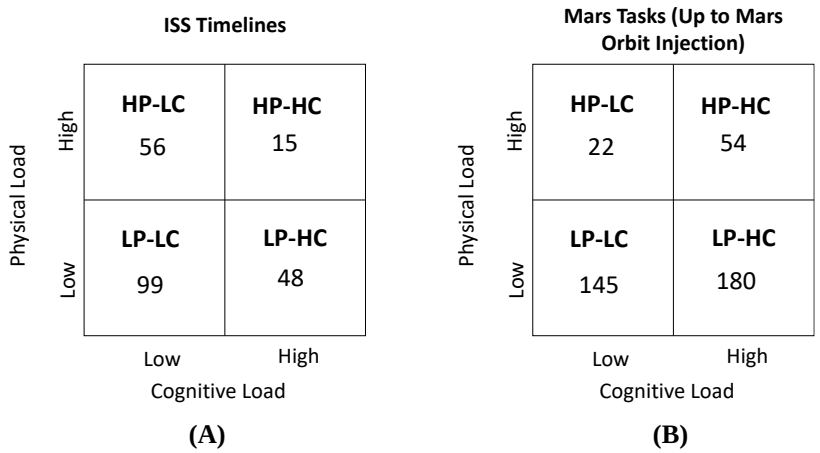
The first step was to identify and downselect representative spaceflight tasks. We used a systematic approach by choosing two sources: the ISS timelines from years of 2000-2014[9] and the Mars Preliminary Task List Report[10]. Our rationale for using the ISS timelines was to provide an accurate and operationally relevant list of tasks that have been conducted inflight, while the Mars Preliminary Task List provided a futuristic look at tasks that are expected for longer duration missions further from Earth.

Due to the number of ISS timelines from 2000-2014, only three years were selected for review: 2000, 2008, and 2014. The intention is that these tasks would be representative of different ISS operational periods from the initialization and operations to full science operation and usage to near end-of-life maintenance and sustainment.

Each of the tasks listed in the timeline was recorded with high level language using verbs to connotate the action involved with the task and the object on which the action was being performed. This was accomplished manually, and only unique tasks were added to the list.

The Mars Preliminary Task List was documented with a similar process. In this case, some similar tasks were defined as unique because of the phase of flight or the specific group to which the task was being conducted differed (for example, communicating with other crewmembers versus communicating with mission control). Because the task list spanned a full Mars mission, the assumption was that the tasks would be symmetric across the last third of the mission, during Earth return. Additionally, we focused specifically on in-flight tasks and therefore, analyzed the task list up to Phase 4 of Mars Orbit Injection. The total number of tasks identified was 619 where 218 tasks were from the ISS timelines and 401 tasks were from the Mars Task List.

With the large number of tasks to select from, we further simplified the process by splitting the task classification into four groups, as shown in Figure 2, and defined as Low Physical–Low Cognitive (LP-LC); Low Physical–High Cognitive (LP-HC); High Physical–Low Cognitive (HP-LC); High Physical–High Cognitive (HP-HC). With this classification, we next categorized each of the tasks from our Total Task List as to whether they demanded high or low physical activity and high or low cognitive activity.



(A) **(B)**

Figure 2. Classification of Task Types. A) Total tasks identified for each Task Type from the ISS timelines. B) Total tasks identified for each Task Type from the Mars Preliminary Task List Report (up to Mars Orbit injection)

We found that while this was a good starting point for characterizing the task types, it was difficult to downselect the tasks further without increasing the resolution of the task characterization. Specifically, since this was considered a feasibility test to determine whether we could identify any differences between tasks as captured by biosignal data, we decided to focus on tasks on the edges of these four quadrants (i.e., the highest physical demand, highest cognitive demand, and lowest physical demand, lowest cognitive demand).

Additionally, when reviewing the tasks, we noted complex nuances to their ranking with regard to potential psychological demand levels of the task. For example, a press conference for the astronauts in space may not require much cognitive demand, but it may be psychologically demanding and therefore perceived as a much more demanding cognitive task. Another example would be a medical emergency in which the physical and cognitive demand may be low, for example to build a makeshift splint for a broken arm. However, the fact that you may be doing this for a friend or crewmate while they are in severe pain may impact your ability to do the task effectively. Therefore, we went back through the list to provide a ranking of the anticipated psychological demand level using values of 1 for low to 5 for high. The subjective ranking was performed by the authors separately and then reviewed collectively to discuss any conflicting results and determine rankings.

Once the psychological rankings were added, we combined the two task lists, removed any non-unique tasks, and selected tasks that were ranked on the ends of physical and cognitive scales, while being low on psychological demand. We were left with the following number of tasks for each of the quadrants: LP-LC (1,1): 36 tasks, LP-HC (1,5): 17 tasks, HP-LC (5,1): 15 tasks, HP-HC (5,5): 19 tasks.

The number of tasks were still too high to effectively measure in the lab, so we removed any experiments that we deemed risky for health and safety reasons or due to lack of resources. Additionally, we removed any tasks that were considered to be too short in duration or would be difficult to implement in the lab setting and would not result in sufficient biosignal data capture. An example of tasks that we removed included: “Clean fingernail”, “Drain sinus passageway”, and “Sleep”. Table 1 shows the final set of downselected tasks.

We next gave this list of 38 tasks to four subject matter experts, including a former astronaut and two NASA employees who work or have worked with astronauts, to solicit their subjective rankings as an initial confidence check and to help further downselect to four optimal tasks we could replicate in the lab for each of the quadrants. With their inputs, along with additional removal of high psychological impact tasks, we decided on the following four representative tasks as a starting point for the pilot study: 1) a basic data entry task as the low physical-low cognitive activity, 2) a timed arithmetic test for the low physical-high cognitive task, 3) running laps for the high physical-low cognitive task, and 4) combined arithmetic and running for the high physical-high cognitive task.

Table 1. List of downselected tasks.

LP-LC	LP-HC	HP-LC	HP-HC
Inventory internal light systems (# of bulbs, light output, etc.)	Perform fluid physics experiment	Test Dynamometer (hand strength measurement)	Egress airlock as part of EVA (in space suit)
Download Data	Data analysis	Stow equipment	Ingress airlock as part of EVA (in space suit)
Measure habitat air quality using Gas Chromatography Sampler (GCS)	Read space crew support system protocols	Perform IRAD (resistive or weightlifting) exercise	Fix a crewmate's dislocated shoulder
Conduct communication checks	Write Conference Report	Perform exercise on stationary bike	Implement countermeasures (e.g., counseling, Software)
Verify controls and switches match checklist	Diagnose and treat sick crewmate	Perform exercise on treadmill	Conduct EVA (moving across habitat in suit)
Monitor displays	Measure pulse, respiratory rate and conduct chest exam (on self)	Egress crew habitat (without spacesuit)	Respond to spacecraft CO_2 alarm
Monitor and control external video camera	Play chess (and other games) with MCC personnel and others on Earth	Ingress to crew habitat (without spacesuit)	Respond to spacecraft general ECLS failure alarm
Conduct simple physical tests (e.g., put pegs in pegboard)	Plan operations (e.g., EVA, maintenance of external components)	Don pressure suit	Respond to spacecraft navigation alarm
Attach and replace sleeping bags and/or tools	Conduct cognitive tests	Respond to spacecraft fire alarm	
Conduct and record solar, planetary and stellar observations	Use sextant and star charts for navigation	Place PHA QDM (Quick Don Mask) on during Emergency (time trials)	
		Perform cardiopulmonary resuscitation (CPR)	

B. Experiment Sensor Suite Selection Process

Selecting a NiBS suite entailed several considerations for the project including the following: what we want to measure, what exists to make the measurement, and what are our constraints in procuring and using the sensor(s). While there are clearly many more constraints and requirements if these systems were to be flown in space as documented by Seyedmadani[11], this project is focused on laying the foundation for the science outputs in a more controlled setting and will be conducting studies in NASA's Human Exploration Research Analog (HERA) as a final test case. With that in mind, the constraints we had for this project are listed as follows:

1. Operational requirements to accommodate use in HERA. Because we would be doing testing at NASA's HERA site, we wanted to use hardware that could be used in their facility. This included the following constraints:
    a. Limited access to internet. Since the analog is meant to mimic spaceflight, there is limited internet access and therefore the hardware must not need real-time connection to the internet to work. This requirement also implies that the system should be able to store its onboard data to be downloaded at the end of the day or task or even at the end of the mission if possible.
    b. Minimal crew time. The lack of crew time drives requirements for the system to be easy to use, easy to setup, and minimal maintenance of the system such as charging or cleaning of the components.

c. Must be mobile. The crewmembers must be able to move around freely while wearing the sensors. Therefore, the hardware cannot be hardwired to a power outlet or to a Data Acquisition (DAQ) system that is not mobile.
d. Compatible with previous systems used in HERA. If the sensor has some history of being used in space or in HERA, this would greatly enhance the ability to compare to previous studies, but this is not a driving requirement.

2. Accessibility of the raw data. The data had to be in an accessible format or in an easily convertible format (.csv, .txt, .xls, .fit) for our team to do post processing and analysis.
3. Low cost/affordability. We were limited to a small hardware budget in our grant and had to have multiple sets because each crewmember in the HERA mission would require their own set. Therefore, we limited our total sensor suite budget to a maximum of $5K. This narrowed the number of systems that could be feasible for our testing.
4. Published results. The system should have some published data, results, or comparative studies against other systems. We wanted a proven and complete solution without having to do major verification tests.

These considerations were combined into a trade study matrix shown in Table 2.

Table 2. Sensor Suite selection considerations. Hardware highlighted in bold were the final selected suite. HR = heart rate, EDA = electrodermal activity, ST = skin temperature, BVP = blood volume pulse, Acc. = acceleration, IBI, = interbeat interval, fNIRS = functional Near-Infrared Sensor, ECG = Electrocardiogram, HbO_2 = forehead oxygenated hemoglobin, HbR = forehead deoxygenated hemoglobin, BP = blood pressure.

Biosignal Measure	What insight does the measure provide?	Hardware that can make the measurement?	Low Cost?	Data Format?	Battery Life	Need wired DAQ?
EDA	Stress, arousal, anxiety	**Empatica E4**	Yes	.csv	24-30 hrs	No
EEG	Activation of certain brain regions	Emotiv; B-Alert	No; No	Need EmotivPro software to convert to .csv; need AcqKnowledge to convert to .csv	4 hrs; 8 hrs	No
Pupil Dilation	Stress, arousal	Tobii Eye Tracking Glasses	No	Need Tobii Pro Lab to convert to .csv	8-9.5 hrs	No
HRV	Workload	**Empatica E4**; BIOPAC; **Biosignalsplux**	Yes; No; Yes	.csv; need BioPac AcqKnowledge to convert to .csv;.txt	24-30 hrs; 24 hrs; 16 hrs	No; Yes; Local DAQ
ECG	Cardiovascular health	BIOPAC; **Biosignalsplux;** Apple Watch	No; Yes; Yes	Need BioPac AcqKnowledge to convert to .csv; .txt; .xml	24 hrs; 16 hrs;18 hrs	Yes; Local DAQ; No
HR	Complement to Workload	**Empatica E4;** BIOPAC; **Polar Chest Strap;** Apple Watch	No; Yes; Yes; Yes	Need AcqKnowledge to convert to .csv; .csv; .xml	24 hrs; ~400 hrs; 18 hrs	Yes; No; No; No
Blood O_2 Level	Absorbed O2 in the blood stream	Pulse Oximeter; Apple Watch	Yes; Yes	Only real-time data; .xml	20-160 hrs (depend on brand); 18 hrs	No; No
BVP	Volume moving through veins	**Empatica E4**	Yes	.csv	24-30 hrs	No
EMG	Activity level of muscles	BIOPAC (BN-EMG2); Moxy	No; Yes	Need BioPac AcqKnowledge to convert to .csv; .csv	24 hrs; 3 hrs	Yes
ST	Stress, arousal	**Empatica E4**	Yes	.csv	24-30 hrs	No
Acc.	Activity and movement	**Empatica E4**; Phillips Actiwatch; Apple Watch	Yes; Yes; Yes	.csv; need Actiware to convert to .csv; .xml	24-30 hrs; 30 days; 18 hrs	No; No; No
fNIRS	Brain activity (via HbO_2 and HbR)	**Biosignalsplux**	Yes	.txt	16 hrs	Local DAQ
BP	Stress or physical activity	**Blood Pressure Cuff**	Yes	.csv or real-time (depend on brand)	Depend on brand	Depend on brand

The suite of sensors listed here is by no means a comprehensive list of all sensors that exist on the market; rather it is based on several desirable characteristics critical to the design of our study. Foremost, the products listed have been favorably reported on by other researchers in this field in the literature and/or at conferences. Additionally, the manufacturers of the highlighted sensor systems provide sufficient information on their website or in their publications to explain system functionality. The remaining trade space included size, portability, commercial availability, and ease of use in an operational setting. With the limitation of cost, we selected the suite of hardware that would provide the most measures for the least amount of hardware.

The downselected sensors are highlighted in bold font in Table 2. The chosen sensors measure a range of parameters most notably consisting of electrodermal activity (EDA) using an Empatica E4 wristband, heart rate (HR) using a Polar H10 chest strap, changes in oxygenated hemoglobin and deoxygenated hemoglobin in the frontal lobe using a portable fNIRS sensor from Biosignalsplux, and an electrocardiogram (ECG) sensor from Biosignalsplux.

C. Design of Experimental Test Session

1. Test Session

We designed a two-hour test session to include the four selected tasks to be performed at the University of Colorado Boulder's Bioastronautics Lab. The session included an introduction during which the subjects reviewed and signed a consent form, put sensors on, and performed control tasks (i.e., sat for three minutes, stood in place for three minutes, and ran in the lab for three minutes). Next, we had them fill out a questionnaire about their background and demographics along with their physical fitness level and food and water intake for the day. We took their seated blood pressure using a blood pressure cuff after each change of activity.

After completing the questionnaire, blood pressure was measured, and then they began the prescribed series of task types starting with the data entry task in which the subject was asked to copy files on a computer from one folder location to another and then rename the file folder. The subject was provided a paper handout of the name of the original files and their corresponding new name and asked to do this task as many times as possible until the test operator told them to stop.

The data download task was followed by running laps back and forth along one length of the ~29-foot lab. The subject was asked to run the laps as fast as they could until the test operator told them to stop after 5 minutes. The subject was not told the duration of the task, as to not allow the subject to pace themselves but rather expend as much energy as quickly as they could to get their heart rate up.

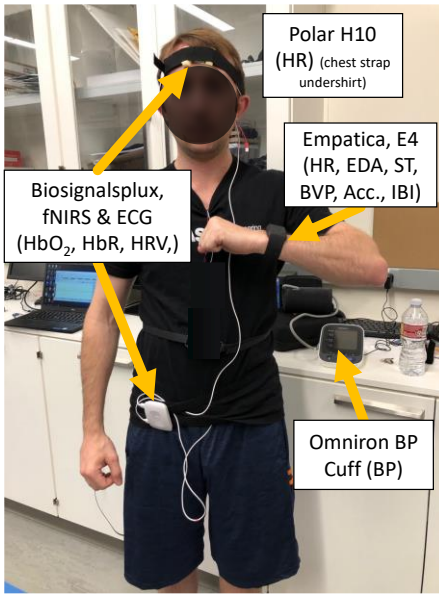


Figure 3. Sensors worn and used during testing.

Following the running task, the subject was asked to sit down and do as many double to single digit arithmetic problems as fast and as accurately as possible. The arithmetic problems were presented on a webpage using a laptop. The subject had to type in their answer using the keyboard for each question.

After the arithmetic task, the subject then did a combined lap running and arithmetic task. This involved the subject starting the arithmetic program on the computer, then running two laps and stopping to answer an arithmetic problem, then the next problem appears and the subject runs two more laps. The subject was again asked to do this as quickly and accurately as possible.

Once the subjects finished these tasks, they were asked to fill out a final questionnaire about their test experience, including how they felt about the sensors, clarity of instructions, and if they had any additional feedback to help improve our test protocol for future subjects.

Each task was to be performed for five minutes with a five-minute break in between. This duration was selected in response to various articles[12,13] showing that this is sufficient to measure physiological responses to high cognitive or physical loads, while minimizing test fatigue or boredom. After each task type, we had the subject take their blood pressure. For the running tasks, if the subject's heart rate surpassed 190 bpm, we'd asked them to stop prematurely if it hadn't been five minutes yet, and if the subject felt the need to stop on their own, they were allowed to end the task. Figure 4 shows the timeline of the test session. We had a total of six participants performing the test session. These four task types were performed with subjects donning a suite of wearable sensors, as shown in Figure 3.

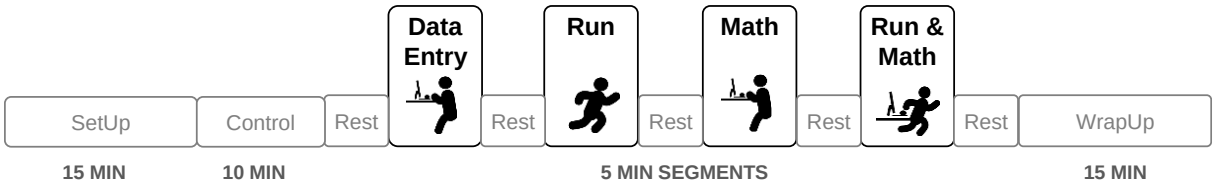


Figure 4. Timeline of test session.

2. Independent Variables

The parameters of interest for this work are two layered. The first layer consists of the basic biosignal data of the test subject throughout the session from the wearable sensors and their performance on the tasks. The second layer of data is the derived function for relating the performance on the task the individual's operational state change. For each of the task types being performed, specific measures of performance were tracked as shown in Table 3.

Table 3. List of performance measures for each task type collected.

Task 1 (LP-LC) ***Data Entry Task***	**TASK 2 (LP-HC)** ***Math Task***	**TASK 3 (HP-LC)** ***Run Task***	**TASK 4 (HP-HC)** ***Run & Math Task***
•Number of Files Transferred •Number of Errors •Total Time Spent on Task •Accelerometer Profile •Energy (derived from Accelerometer) •Peak Acceleration •Min Acceleration	•Number of Questions Answered •Number of Errors •% Correct •Total Time Spent on Task •Accelerometer Profile •Energy (derived from Accelerometer) •Peak Acceleration •Min Acceleration	•Number of Laps •Pace per Lap •Total Time Spent on Task •Accelerometer Profile •Energy (derived from Accelerometer) •Peak Acceleration •Min Acceleration	•Number of Laps •Pace per Lap •Number of Questions Answered •Number of Errors •% Correct •Total Time Spent on Task •Accelerometer Profile •Energy (derived from Accelerometer) •Peak Acceleration •Min Acceleration

IV. Preliminary Results

To tune the test sequence, each pilot scenario was performed with slight permutations in reaction to observations and feedback from previous tests. While all subjects completed all four task types, no sequence was precisely repeated. The purpose of these varied test scenarios was three-fold: 1) to guide decisions regarding the test session setup 2) to guide insights into the data collection process and 3) to guide decisions regarding the type of post processing that might be applicable.

A. Insights on Test Setup

This preliminary work led to several important realizations for our test setup. Specifically, we were able to identify hardware and software issues that could be fixed to improve the data quality and maximize our ability to compare between the biosignals and task types.

Hardware issues that arose included difficulty with strapping the Biosignalsplux DAQ in a consistent and safe place around the subject as they ran laps. If the subject had pockets, we would put it in their pocket, or have it clipped to the elastic band on their pants or shorts. But this placement was not conducive to good running form with subject feedback suggesting they worried it might fall out or off. Additionally, the wires from the ECG and fNIRS also deterred the subject from attempting a normal running gait as the wires often would swing across their face or dangle by their hand. A better harnessing system to tightly contain the wiring and the DAQ would serve to minimize the subject's ability to get tangled with the wiring and accommodate their gait to protect the hardware.

In terms of the length of the task, completing all four task types with breaks in between took about two hours for each participant. This was concerning as we were worried about an order effect where fatigue could become a confounding effect on the later tasks. Additionally, the placement of the forehead fNIRS becomes uncomfortable and irritating after about an hour so we consequently decided to reduce the number of task types to test going forward.

The fNIRS forehead sensor was deemed very uncomfortable by many of the subjects, to the point that they had to re-adjust the band during the testing. The sensor design includes two LEDs that protrude about 5 millimeters from the base. These LEDS are then pressed into the forehead with an elastic band around the head. The tightness of the band does seem to affect the data outputs and will be further investigated prior to deploying a full data collection regime.

Another consideration for future improvements for the test setup includes ensuring the subject wears appropriate attire for physical activity. The testing was often done when the subject had a break in between classes and therefore the subject would be wearing casual attire rather than active wear. This was an issue for some of the subjects who seemed more constrained in their running gait.

Also due to the size of the lab space, each lap was only 29 feet in length and to run fast, the subject had to abruptly stop on one end and quickly spin around to run to the other side. This abrupt stopping was not conducive to good running form and further exacerbated motion artifact. An ideal setting may have been outdoors on a track or on a treadmill.

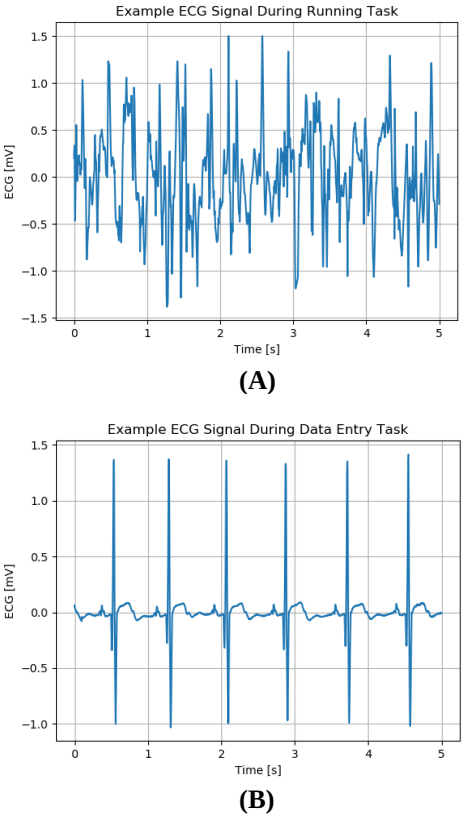


Figure 5. Comparison of a sample ECG signal (A) during the running task with large motion artifacts vs (B) seated data entry task.

B. Insights on Data Collection Process

Collecting the correct data and collecting it accurately and consistently was also something we reviewed from this pilot test. To verify the correct data was being collected, we compared the biosignal data ranges to existing literature regarding expected values. For example, baseline healthy heart rate should be in the range of 60-100 bpm, while highly physical tasks should result in an increased heart rate usually around 128 bpm to 170 bpm, but no more than 220 bpm minus the age of the subject[14].

To increase accuracy and consistency of the data collection there were several updates to the protocol that had to be made. First, the placement of the sensors could be better standardized. With the forehead fNIRS, location of the sensor is important as specific parts of the frontal lobe will be activated for certain tasks[15].

Second, the data showed that the running task created a high degree of motion, producing large amounts of spurious data in the ECG plots. This large increase in spurious results would likely make it impossible to extract data on heart rate variability and led to our decision to switch the running task to a stationary biking task for later tests. Third, we realized the task types could be considered more on a continuous spectrum, where even the rest in between tasks could be considered as a low physical low cognitive task.

The testing sequence is susceptible to a variety of confounding factors which may have dependency on ordering of tasks within the sequence. For example, long periods between tasks, repetition of certain tasks, or insertion of differing tasks between repeated tasks may alter the subject's performance despite the task being otherwise equivalent.

C. Insights on Data Analysis

From this preliminary engineering data, several insights could be gleaned on how to maximize comparative power between measures of the biosignals, the task types, and individual differences.

Looking at the results, it became apparent that there could be a variety of approaches to sub-sampling the data in time. Due to the differing task lengths, some data streams would only be available for three minutes while others

would be available for five minutes. To account for such variations, the analysis would need to be tailored to each task type, adding a new layer of complexity to developing generalizable predictions across varied task types. The analysis must also determine optimal window sizes and window location for comparing biosignal data. For derived data such as HRV, the window size of the data being analyzed affects the mean values. Additionally, the location at which the signal is measured is important as well. In the case of heart rate, some subjects did not see an immediate rise in their heart rate during the running tasks as they were slow to start the task (i.e., started with a slow jog before getting into a more steady and faster rhythm) or there might be an association to the individual's fitness level, in which the higher the fitness level, the less immediately impactful the running task is to their heart rate. The average HR would look different if the measure was over the entire five-minute window as compared to just a 30 second window at the end of the task.

Beside considering what window size and locations for the data analysis, other considerations that appeared included how to compare the cohort of subjects and extract individual differences. Such an approach would need to isolate individual differences or at least identify grouping of subjects for analysis. One grouping may involve fitness levels or another may select subjects with similar skill level in arithmetic. Other groupings could be based on physiological considerations such as height or body mass index (BMI) or even baseline biosignal measures.

Another realization after looking at this initial data was that the five-minute break in-between sessions may not be sufficient for 'resetting' the subject to a baseline state, especially after a high physical activity. Ordering of the tasks may also impact the overall performance as the baseline of the subject gets adapted with respect to the previous task.

The variation in task time added a further wrinkle that direct comparison of different task types becomes clouded. This is because the number of cycles within the given task is not consistent, thereby changing the relative strength of the analysis statistics. While such a finding could be considered obvious, it presents an interesting analysis point as real tasks will occur over widely varying time durations. This again points back to the generalizability concern noted previously.

Another big challenge in the analysis stems from time offset between each sensor. The various sensors are not part of a common platform and therefore it is not possible to start each system simultaneously. This meant that, for example, the Polar chest strap ran on an internal clock which could be referenced to a starting time differently than the Biosignalsplux. Fortunately, these hardware devices can be set to output time in standard UTC. So even though the starting point and relative timing between the devices might be offset, the time could be shifted in post-processing to be consistent. This does however add additional analysis steps and is not easily automated. Ultimately, the pilot testing illuminated a range of both procedure and analysis hurdles that were then factored into the subsequent testing protocols. It is noted that an updated test protocol for the upcoming testing regime was submitted and approved under the University of Colorado Boulder's IRB under the number 20-0003.

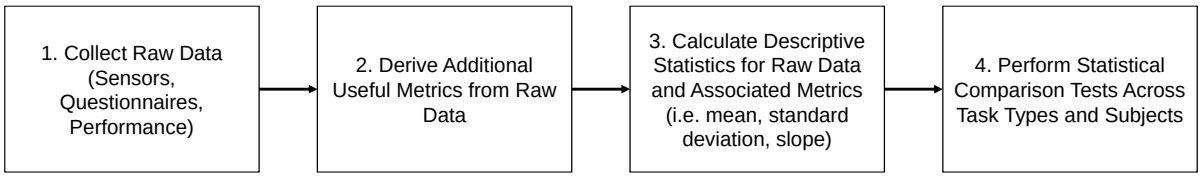


Figure 6. Flow of data processing and analysis.

V. Discussion and Conclusion

With this initial work, we have developed a standard method to generate different representative mission relevant task protocols, procured a suite of biosignal sensors, analyzed early pilot data, and developed data analysis protocols for integrating the data products into our modified version of a crew performance model. The insights garnered from the pilot testing were used to develop a new protocol to allow for conducting 'at home' remote research in light of the lab access restrictions incurred during the ongoing COVID-19 pandemic. We demonstrated feasibility of our approach and are establishing the next phase of this project where we will begin a larger data collection effort and analyze the data for assessing predictive patterns that can be applied toward ultimately validating our task design framework.

This process highlights that the conceptual model proposed by Fanchiang[7] requires continual updates which are dependent on both the type of data that can be non-invasively captured to describe changes to crewmember capabilities and on what sensors exist to feasibly make the measurements. While the tasks presented are clearly simplistic

representations compared to actual spaceflight operations, this initial pilot study allowed us to systematically look at the tangible measurements and determine the relationship between the biosignals and expected performance on a task.

One potential application of this method is as a proxy for measuring the quality of a particular task design. Such a methodology is specifically relevant for scenarios in which humans will rely significantly on automation and robotic interfaces in performing various tasks. While models have been developed to use physiological measures as a proxy for cognitive or physical load of a particular task, few have used the measures to evaluate and adjust the task using knowledge of the incoming operational state of the human.

Our approach specifically frames the design of the task around the dynamic changes to the crewmember's operational state. This viewpoint adds a layer of operational relevancy to the current state of human-robot interactions where instead of just asking "how much cognitive load does this task require?", this framework can start to ask more dynamic questions such as "if this human is in a particular operational state are they able to perform the upcoming task", and/or "what type of performance could be expected on this task give the incoming operational state of the human".

This work has helped drive improvements for future human subject testing regimes that will be conducted from 2021-2023, and it has provided useful insight for the type of data analysis and processes needed for full operational scenarios in NASA's HERA missions.

Acknowledgments

We would like to thank Dr. Steve Robinson at the University of California Davis for organizing the multiple efforts involved with the HCAAM VNSCOR Grant 80NSSC19K0655, and Dr. Brian Gore at NASA Ames for his help with managing our grant. We would also like to acknowledge additional support provided by the William F Marlar Foundation for equipment purchases.

References

[1]Fanchiang, C., Klaus, D. M., Gore, B. F., Marquez, J. J. Survey and Assessment of Crew Performance Evaluation Methods Applicable to Human Spacecraft Design. IEEE Aerospace Conference 2015. Big Sky, Montana. Mar 1-8, 2015.

[2]Keenan, A., Young, M., Saile, L., Boley, L., Walton, M., and Kerstman, E. (2015). "The Integrated Medical Model: A probabilistic simulation model predicting in-flight medical risks," in 45th International Conference on Environmental Systems, 2015.

[3]Wilson GF, Russell CA. (2003). Operator functional state classification using multiple psychophysiological features in an air traffic control task. Hum Factors. 2003 Fall;45(3):381-9. doi: 10.1518/hfes.45.3.381.27252. PMID: 14702990.

[4]Wilson GF, Russell CA. (2003). Real-time assessment of mental workload using psychophysiological measures and artificial neural networks. Hum Factors. 2003 winter; 45(4):635-43. doi: 10.1518/hfes.45.4.635.27088. PMID: 15055460.

[5]Hockey, G.R.J., Gaillard, A.W.K., Burov, O.(eds) (2003). Operator Functional State: The assessment and Prediction of Human Performance Degradation in Complex Tasks Vol. 355 of NATO Science Series, I: Life and Behavioural Sciences ISBN: 978-1-58603-362-0.

[6]Lohani, M., Payne, B. R., and Strayer, D. L. (2019). *A Review of Psychophysiological Measures to Assess Cognitive States in Real-World Driving* Front. Hum. Neurosci., 19 March 2019 | https://doi.org/10.3389/fnhum.2019.00057.

[7]Fanchiang, C. 2017. *A Quantitative Human Spacecraft Design Evaluation Model for Assessing Crew Accommodation and Utilization.* Boulder, CO: Doctoral Thesis. University of Colorado Boulder.

[8]Fanchiang, C., Marquez, J.J. and Klaus, D.M. (2020) *A Framework for Relating Crew Member Performance Measures to Spacecraft Design and Operations*. New Space 8(4): 193-200 DOI: 10.1089/space.2020.0033

[9]NASA. *International Space Station Timelines.* Accessed March 3, 2021. https://www.nasa.gov/mission_pages/station/timelines/index.html.

[10]Stuster, J. W., J. A. Adolf, V. E. Byrne, and M. and Greene. (2018). *Human Exploration of Mars: Preliminary Lists of Crew Tasks.* NASA. NASA/CR-2018-220043.

[11]Seyedmadani, K. Gavalas, L. Tucker, K. (2019) "Development of spaceflight biomedical and health hardware according to the needs of users in medical device and diagnostic industry." NASA ISSMP, ASMA, May 2019.

[12]Pattyn, N., Neyt , X., Henderickx, D., Soetens, E. (2007) *Psychophysiological investigation of vigilance decrement: Boredom or cognitive fatigue?* Physiology & Behavior 93 (2008) 369–378.

[13]Head JR, Tenan MS, Tweedell AJ, Price TF, LaFiandra ME and Helton WS (2016) *Cognitive Fatigue Influences Time-On-Task during Bodyweight Resistance Training Exercise.* Front. Physiol. 7:373. doi: 10.3389/fphys.2016.00373.

[14]American Heart Association. Target Heart Rates. Accessed: 8 March 2021. https://www.heart.org/en/healthy-living/fitness/fitness-basics/target-heart-rates.

[15]Tanidaa, M., Sakatanib, K., Takanoc, R., and Tagaic, K. (2004). *Relation between asymmetry of prefrontal cortex activities and the autonomic nervous system during a mental arithmetic task: near infrared spectroscopy study*. Neuroscience Letters 369 (2004) 69–74.