Evaluation of the Reliability and Criterion Validity of the Performance Assessment Tool

BAA AFRL-RQKHC-2016-0009 Aerospace Medicine, Clinical Research, Human Systems Integration Research, and Expeditionary Medicine

—Submitted by—

The Florida Institute for Human and Machine Cognition (IHMC)
40 S. Alcaniz Street
Pensacola, FL 32502
850-202-4400

DUNS: 158995759   Cage Code: 312T7   IHMC Business Size: 110   NAICS: 541715 Business size 1000

Administrative Contact
Diana Thacker
Director of Sponsored Research
FSO
(Address classified to Diana Thacker
at IHMC address above)
dthacker@ihmc.us
850-202-4473
Fax: 850-202-4440

Technical Contact
Jeff Philips, Ph.D.
Research Scientist
jphillips@ihmc.us
850-202-4460
Fax: 850-202-4440

Background

The United States Air Force School of Aerospace Medicine is leading a series of studies to evaluate the effects of operational stressors on human cognitive and perceptual performance. The goal of these studies is to better understand and quantify the effects of operational stress on human performance and develop effective mitigation strategies. Common occupational stressors that have been identified to significantly degrade human cognitive and perceptual performance include: hypoxia, hypocapnia, thermal stress, fatigue, motion sickness, spatial disorientation, and high cognitive workload. Unfortunately, many psychometric tools used to identify hypoxia-related performance deficits are limited in their utility due to two issues. First, these measures possess low test-retest reliability. Low test-retest reliability results in high error variance, which makes detection of hypoxia-related performance deficits improbable (Kerlinger, 1973). Second, classic psychometric tools that do detect hypoxic stress reliably (i.e., Simple Reaction Time, Choice Reaction Time) lack face validity and are simplistic when compared to relevant operational tasks. A lack of operational face validity leads to the argument among operators and military leadership that indicators of hypoxia identified through these classic metrics of human performance are irrelevant to real-world, operational settings (Nevo, 1985).

In an attempt to develop a cognitive-perceptual task that is sensitive to the effects of stressors while also having high operational face validity, the United States Air Force School of Aviation Medicine in cooperation with the University of Notre Dame developed the Performance Assessment Tool (PAT). The PAT is a computerized, multi-tasking performance assessment involving one primary task of tracking and three additional tasks including math, memory, and mannequin (see PAT subsection below for further explanation). Importantly, the PAT requires similar cognitive processes as flying. Initial validation studies revealed that the PAT possesses high test-retest reliability, however the task composition has changed significantly since those studies were conducted.

The current study was performed to address two specific aims: 1) To provide a second evaluation of PAT's reliability across twenty sessions conducted over two days, and 2) To determine PAT's sensitivity to the performance decrements associated with the stressor of hypoxia. In other words, this study is evaluating PAT's criterion validity to identify hypoxia compared to another measure, CogScreen Visual Sequence Comparison, that is a "gold standard" for identifying hypoxia but lacks face validity.

Method

PAT

Often, pilots are thought to have three main tasks to fly: 1) aviate, 2) navigate, and 3) communicate. The PAT was designed to be akin to these tasks relevant for aviation. The PAT is performed on a computer by looking at a monitor and controlling inputs using a joystick or computer mouse; we used a joystick in our study and will continue to refer to the use of a joystick instead of a mouse. Visual representations of the PAT are in Figures 1, 2, and 3.

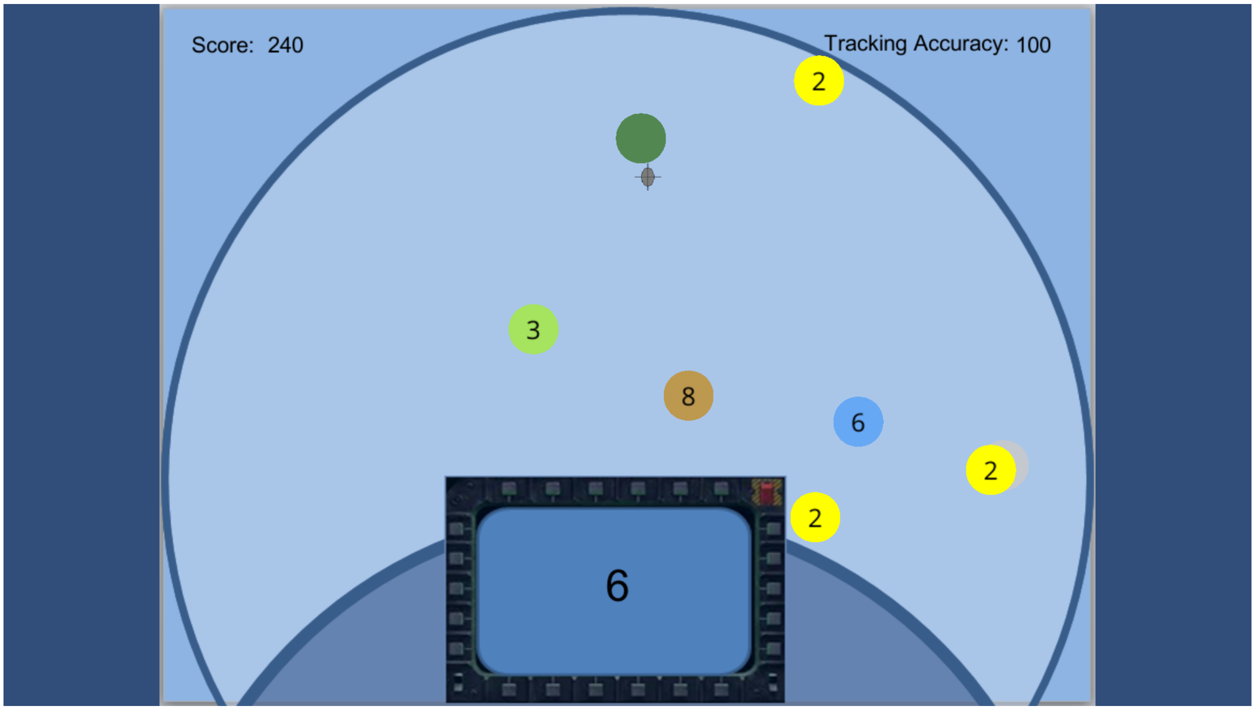

Figure 1. Visual depiction of the PAT during a math task presentation.

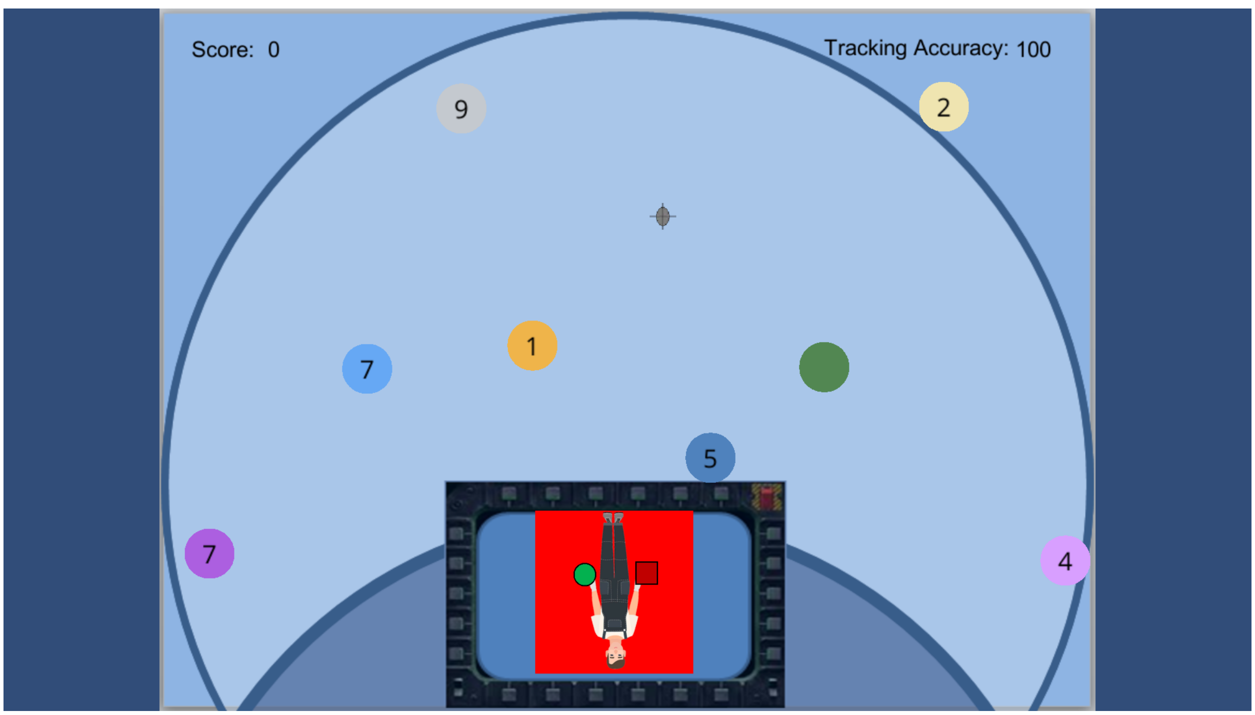

Figure 2. Visual depiction of the PAT during a mannequin task presentation.

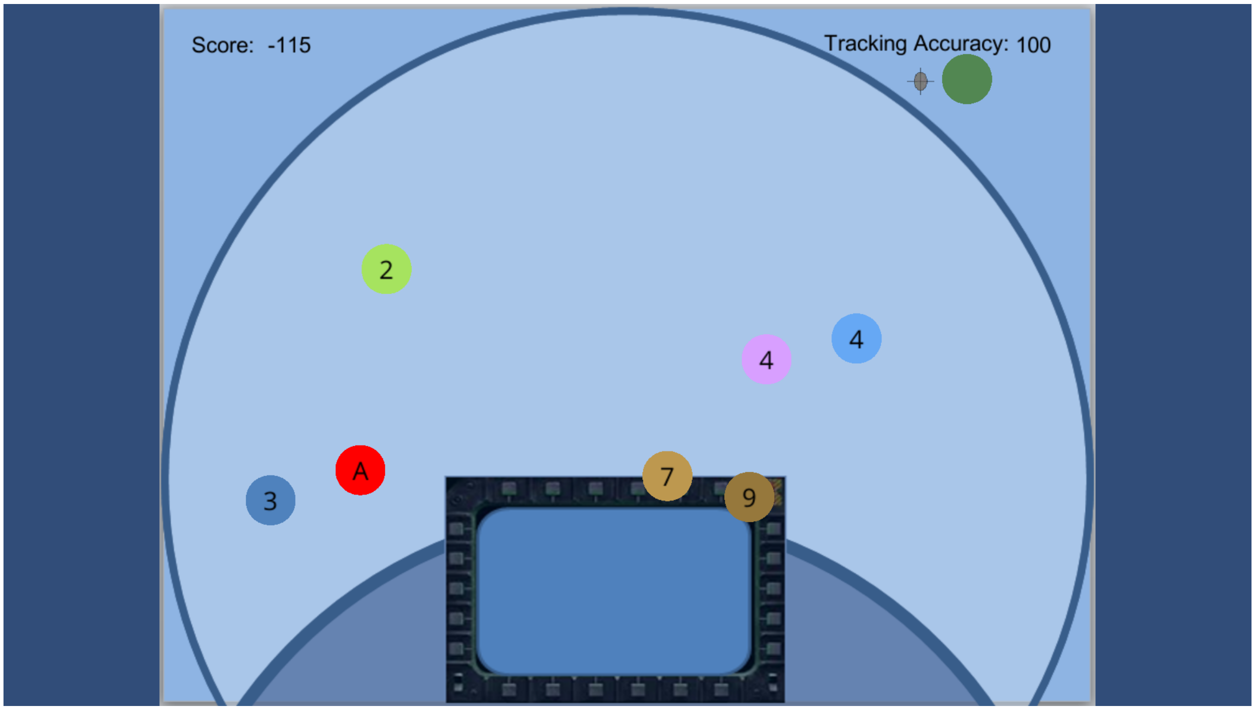

Figure 3. Visual depiction of the PAT during a memory task presentation.

The primary task of the PAT is to maintain tracking, which involves using the joystick to control a cursor on the screen to track the solid green circle that moves at random. Distractor circles are presented throughout the PAT, which are circles of different colors with a number inside the circle; the green circle that should be tracked does not have a number inside of it. The tracking task is supposed to be prioritized, as it is intended to be analogous to the aviate aspect of pilot's tasks while flying. In addition to tracking, the PAT involves performing three additional (or secondary) tasks that occur sequentially of each other but simultaneously with tracking; again, this is akin to having to aviate while performing other duties in-flight. Pilots have to perform many tasks in tandem and use working memory to maintain information being presented to them on their controls in order to aviate. Since flight duties require working memory, one of these additional tasks presented is math. For the math task (see Figure 1), three circles with numbers inside of them are presented on the screen. The participant has to sum these numbers together and determine if their sum is equivalent to the value presented in the multi-functional display. This is thought to utilize visual working memory, as participants are temporarily holding the values that they perceived visually in memory, processing these values by summing them, and then responding based upon this sum held temporarily in memory.

One of the other two secondary tasks is mannequin (see Figure 2); this spatial apperception task involves viewing either an aircraft or human that has an object either to the right or left of them, and the aircraft or human is rotated. This is a spatial rotation task in which one has to identify if the object in the right hand or under the right wing, from the pilot's perspective, is the same color and shape as the background. Since the mannequin task requires visual-spatial processing, it is analogous to the navigate aspect of flying.

A final task that the PAT presents is memory, in which a string of letters and numbers are presented auditory. This string has to be held in short-term memory until a red circle appears on the screen with a letter or number. The participant has to determine whether that stimulus was in the string presented previously (see Figure 3). Typically, pilots have to take information being communicated to them, keep it in short-term memory, and then apply that information in some way. This memory task is analogous to the cognitive processes utilized while communicating with air traffic controllers in aviation.

Participants

Twenty-five aeromedically cleared student Naval Aviators at the Naval Air Station Pensacola, FL participated in the study. All were men with ages ranging from 22 to 29 years (M = 23.48 years, SD = 1.56 years). Their flight hours ranged from 0 to 100 (M = 11.16 hours, SD = 23.55 hours). Ninety-six percent of the student aviators were right-handed.

Procedure

The study consisted of three separate sessions that spanned across three days. The purpose of the first and second sessions was threefold: 1) to train participants on the PAT, 2) ensure performance asymptotes prior to the testing under the stressor (i.e., hypoxia), and 3) perform a test-retest reliability assessment. The third session was intended to assess PAT performance under an environmental stressor (i.e., hypoxia) and compare it to CogScreen Visual Sequence Comparison, a task known to be a sensitive measurement of the performance effects of hypoxia exposure.

The first and second sessions (Days 1-2) had the same task manipulations of the PAT (see Table 1). The PAT was first performed in its entirety for two trials at the beginning of each day in order to orient the participant to task. Afterwards, participants were trained gradually on each subtask, starting with tracking only and building up to the full PAT.

Table 1 First and second session (Days 1-2) PAT task manipulations

Trial	Task
1	Full PAT
2	Full PAT
3	Tracking
4	Tracking + Math
5	Tracking + Mannequin
6	Tracking + Memory
7	Tracking + Math + Memory
8	Tracking + Math + Memory
9	Tracking + Math + Mannequin
10	Tracking + Math + Mannequin
11	Tracking + Mannequin + Memory
12	Tracking + Mannequin + Memory
13	Full PAT
14	Full PAT
15	Full PAT
16	Full PAT
17	Full PAT

18	Full PAT
19	Full PAT
20	Full PAT
21	Full PAT
22	Full PAT

The third session (Day 3) was different from the two aforementioned sessions, as this was the session that involved a hypoxia exposure. Prior to beginning the session, participants were equipped with a pulse oximeter, transcutaneous CO2 monitor, a portable EEG system, and an eye tracking system. These devices were included to monitor the health status of the participant and collect psychophysiological data during the hypoxia session. During the third session, participants alternated between performing two-minute trials of the Full PAT, the tracking only portion of the PAT, and CogScreen for a total period of 40 minutes; for all participants, the first 20 minutes was the normoxia exposure and the last 20 minutes was the hypoxia exposure. The initial 20-minute session was completed under normoxia conditions through the Reduced Oxygen Breathing Device (ROBD). Following the normoxia exposure of the experiment, participants were switched to a 10.5% O2 mixture (18,000ft equivalent) using the ROBD for 20 minutes or until the participant's SPO2 dropped below 60%. Presentation order of the full PAT, tracking only, and CogScreen were counterbalanced. Each task (PAT, tracking, CogScreen) was presented was repeated three times in each 20 session (normoxia and hypoxia), resulting in 6 total trial presentations per task (3 for normoxia and 3 for hypoxia). An example of the task trials performed by a participants are presented in Table 2, noting that order of the tasks was counterbalanced across participants.

Table 2 Session 3 Hypoxia Normoxia exposure. Note the order of tasks was counterbalanced.

Trial	Task
	Normoxia Exposure
1	CogScreen
2	Full PAT
3	Tracking
4	CogScreen
5	Full PAT
6	Tracking
7	CogScreen
8	Full PAT
9	Tracking
	Hypoxia Exposure
10	CogScreen
11	Full PAT
12	Tracking
13	CogScreen
14	Full PAT
15	Tracking
16	CogScreen
17	Full PAT
18	Tracking

Experimental Design

Specific Aim 1: Reliability. The final iterations of the full PAT containing all subtasks on Days 1 and 2 were assessed in the reliability analysis. There were 10 final full PAT iterations on Days 1 and 2 (Trials 13-22 in Table 1), making 20 iterations in total that were included in the reliability analysis. Since the goal of this analysis was to determine if the metrics used to define PAT performance measures the same construct reliably and high test-retest reliability, we computed Cronbach's alpha as a measure of reliability or consistency over iterations. As the intercorrelations across iterations increases, Cronbach's alpha will also increase, indicating high reliability. Several metrics of PAT performance were assessed, including performance on tracking, math, mannequin, memory, as well as two composite scores called PAT Composite and

PCOLA Composite. These measures are explained in the Performance Metrics of the PAT subsection below.

Specific Aim 2: Criterion Validity. On Day 3, participants performed three cognitive tasks repeatedly during a normoxia session first and then a hypoxia session that followed (as displayed in Table 3). The cognitive tasks were the full PAT, PAT tracking only (i.e., no multitasking), and CogScreen. CogScreen was selected as the gold standard for comparison with PAT performance under stress due to CogScreen's acceptance by the Federal Aviation Administration as a neurological assessment tool and its documented sensitivity to hypoxic stress (Rice et. al., 2003). The between-subjects variable was stress (no stressor/ normoxia vs. stressor/ hypoxia). The within-subjects variable was trial, such that each cognitive task was performed 3 times during the normoxia session and 3 times during the hypoxia session. A 2 (normoxia vs. hypoxia) x 3 (trial) Factorial Repeated-Measures Analysis of Variance was computed for each performance metric. The performance metrics are explained below.

Performance Metrics of the PAT. The performance assessment tool is newly developed, and as such, questions remain unanswered regarding the way to score the PAT. During the Practice Phase (Day 1) and Acquisition Phase (Day 2), PAT difficulty level for each subtask varied based on the individual's performance; we opted for difficulty level to vary in order to determine the optimal difficulty level for PAT Tracking and the subtasks for each individual to use during the stressed/hypoxia testing (Day 3). Difficulty level for the tracking task is solely dependent upon tracking performance, but difficulty level for the other subtasks is tied to both tracking performance as well as performance on the specific subtask. Thus, if a participant is tracking accurately above a certain threshold and gets three math problems in a row correctly, then the participant's difficulty level for math will increase by one. This further enforces the tracking task as the primary task.

In the standard PAT output, there was not a variable that combined tracking performance with the difficulty level of the tracking task. Therefore, tracking performance for individuals who reached higher difficulty levels could be lower than those operating at a lower difficulty level, as higher difficulty resulted in performance decrements compared to lower difficulty. For the Math, Mannequin, and Memory task scores, the standard PAT output does generate a score that takes into account difficulty level, but the underlying algorithm used to compute these scores is unknown to the researchers.

Therefore, several variables were created that combine performance and difficulty level into one score. The final list of variables of interest that were included in our analysis were based on discussions with the sponsor and the PAT design team, in addition to pilot work conducted prior to data collection. One variable of interest, the PAT Composite, was present in the standard PAT output and was used without further calculation. All other new variables were calculated from variables included in the standard PAT output file. Below is the list of variables of interest analyzed in this report, their definition, and the formulas used to calculate them.

Tracking: This measurement is a measurement of tracking efficiency based on the percent on target per session and the final tracking difficulty level.

$$Tracking = \frac{\%\ On\ Target}{10} \times Final\ Tracking\ Difficulty\ Level$$

Math: This variable is a measurement composed of the total number of correct math items per session and the final math difficulty level.

$$Math = \%\,Correct\ Math\ Items\ \times Math\ Final\ Difficulty\ Level$$

Mannequin: This variable is a measurement composed of the total number of correct mannequin items per session and the final mannequin difficulty level

$$Mannequin = \%\,Correct\ Mannequin\ Items \times Mannequin\ Final\ Difficulty\ Level$$

Memory: This variable is a measurement composed of the total number of correct memory items per session and the final memory difficulty level.

$$Memory = \%\,Correct\ Memory\ Items\ \times Memory\ Final\ Difficulty\ Level$$

Pensacola Composite (PCOLAC): This is a variable that combines Tracking2, Math 1, Mannequin1, and Memory1 into a composite score that also takes final difficulty level into account.

$$PCOLAC = Tracking2 + Math + Mannequin + Memory$$

PAT Composite: As mentioned above, the PAT Composite Score is the standard PAT composite of performance across tracking and the three secondary tasks.

$$PATC = Dual\ Task\ Total\ Score\ (combination\ of\ tracking\ and\ math\ score)$$
$$+ Mannequin\ Total\ Score + Memory\ Total\ Score$$

Results

Specific Aim 1: Reliability Analysis

Cronbach's alpha was calculated (see Table 3) as a measure of reliability for each performance metric defined above across the last ten Day 1 sessions, across the last ten Day 2 Sessions, and across these 20 sessions of Day 1 and Day 2 combined.

Table 3 Cronbach's alpha calculations across the 10 iterations for Days 1 and 2 and 20 iterations for the combination of Days 1 and 2.

Variable	Day 1	Day 2	Day 1 and 2
Tracking	0.674	0.917	0.890
Math	-0.202	0.265	0.411
Mannequin	0.493	0.642	0.729
Memory	0.350	0.440	0.501
PCOLAC	0.239	0.909	0.709
PATCOMPOSITE	0.495	0.733	0.762

The general conventions for Cronbach's alpha are that values between 0.8 and 0.9 possess good reliability and an alpha above 0.9 possesses excellent reliability (Ponterotto & Ruckdeschel, 2007). PAT variables possessed low reliability across the first 10 sessions completed during Day 1. Conversely, the reliability of PAT improved dramatically across the second ten sessions completed on Day 2 with PCOLAC (0.909) and Tracking (0.917) possessing the highest alpha values across Day 2, demonstrating excellent reliability according to convention. The PAT composite showed acceptable reliability across Day 2 with an alpha value of 0.733. Reliability analysis across the 20 sessions from Days 1 and 2 showed good reliability for Tracking (0.890) and acceptable reliability for PCOLAC (0.709).

Together, this demonstrates that PAT possesses high reliability that is more than adequate to justify its use in human performance and stress work. The reliability of an assessment is important, as consistency in an assessment means the within-condition or error variance is minimized, thus making the assessment capable of detecting changes in performance between conditions (i.e., Stressed vs. Unstressed).

Criterion Validity Analysis and Results

To evaluate the sensitivity of PAT to changes in human performance associated with exposure to hypoxia, a series of A 2 (normoxia vs. hypoxia) x 3 (trial) Factorial Repeated-Measures ANOVAs were conducted on each performance metric of interest for the PAT and the CogScreen VSC. Results complete with *p*-values and effect sizes are available on Table 4. The analysis suggested that PAT is very sensitive to measuring the performance effects of hypoxia on psychomotor performance, demonstrating not only significant effects but importantly large effect sizes. Cohen's (1988) conventions for interpreting effect sizes associated with partial eta-squared values are as follows: small=.01, medium=.06, and large=.14. The significant interaction effects associated with the PCOLAC and Math scores showed that these metrics are sensitive to accumulating performance effects during hypoxia associated with time at altitude (see Figures 6 and 9). Among variables of interest, tracking only possessed the largest observed effect size. The next largest observed effect sizes were PCOLAC and PAT Tracking, respectively (see Table 4).

Cogscreen Visual Sequence Comparison, a gold standard for measurement of hypoxic effects, possessed the largest effect size for a main effect of stressor among the measures included in this analysis. Importantly, CogScreen VSC did not possess a significant stressor x time interaction, suggesting that VSC was not sensitive to changes in performance associated with time at altitude (see Figure 11). The PAT composite metric PCOLAC and the Math variable did have significant stressor x time interactions, and several other PAT metrics seemed to show interaction trends (see Table 4).

Table 4 ANOVA Results for each variable main effects, and interaction effects

2 (Normoxia vs Hypoxia) X 3(Time 1, Time 2, and Time 3) Repeated Measures ANOVA Results						
Variable of Interest	Source	Degrees of Freedom	F-Value	P-Value	Partial Eta Square	Correction
Tracking	Stressor	1, 24	18.194*	0.000	0.431	None
	Time	1.154, 27.693	0.864	0.376	0.035	G-G
	Stressor*Time	1.284, 30.810	2.797	0.096	0.104	G-G
PAT Tracking	Stressor	1, 24	15.746*	0.001	0.396	None
	Time	1.62, 38.767	0.708	0.470	0.029	H-F
	Stressor*Time	1.305, 31.325	0.515	0.601	0.021	G-G
Math	Stressor	1, 24	7.836*	0.010	0.247	None
	Time	1.597, 38.326	3.513*	0.049	0.128	H-F
	Stressor*Time	1.465, 36.885	4.465*	0.028	0.157	G-G
Mannequin	Stressor	1, 24	4.853*	0.037	0.168	None
	Time	1.68, 38.218	1.662	0.205	0.065	None
	Stressor*Time	2, 48	1.473	0.239	0.058	H-F
Memory	Stressor	1, 24	4.795*	0.039	0.167	None
	Time	2, 48	1.903	0.160	0.073	None
	Stressor*Time	2, 48	1.407	0.255	0.055	G-G
PCOLAC	Stressor	1, 24	14.630 *	0.001	0.379	None
	Time	1.369, 32.860	4.373 *	0.033	0.154	G-G
	Stressor*Time	1.650, 39.609	4.085 *	0.031	0.145	H-F
PAT COMP	Stressor	1, 24	13.627*	0.001	0.362	None
	Time	2, 48	3.477*	0.039	0.127	None
	Stressor*Time	1.72, 41.26	1.266	0.289	0.050	H-F
Cog Screen VSC	Stressor	1, 21	27.218*	0.000	0.564	None
	Time	2, 42	2.257	0.117	0.097	None
	Stressor*Time	2, 42	3.160	0.053	0.131	None

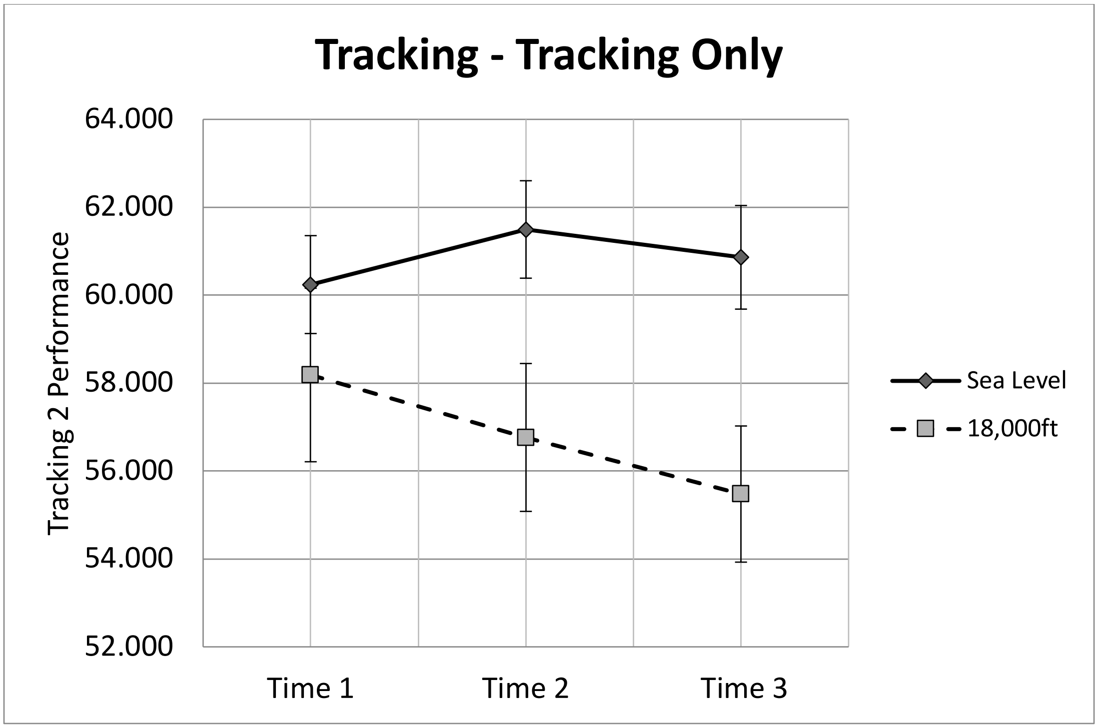

Figure 4: Significant main effect for Stressor $F(2,24) = 18.194$, $p < .001$, $P\eta^2 = .431$. Significant differences between hypoxia and normoxia were found in time points 2 and 3.

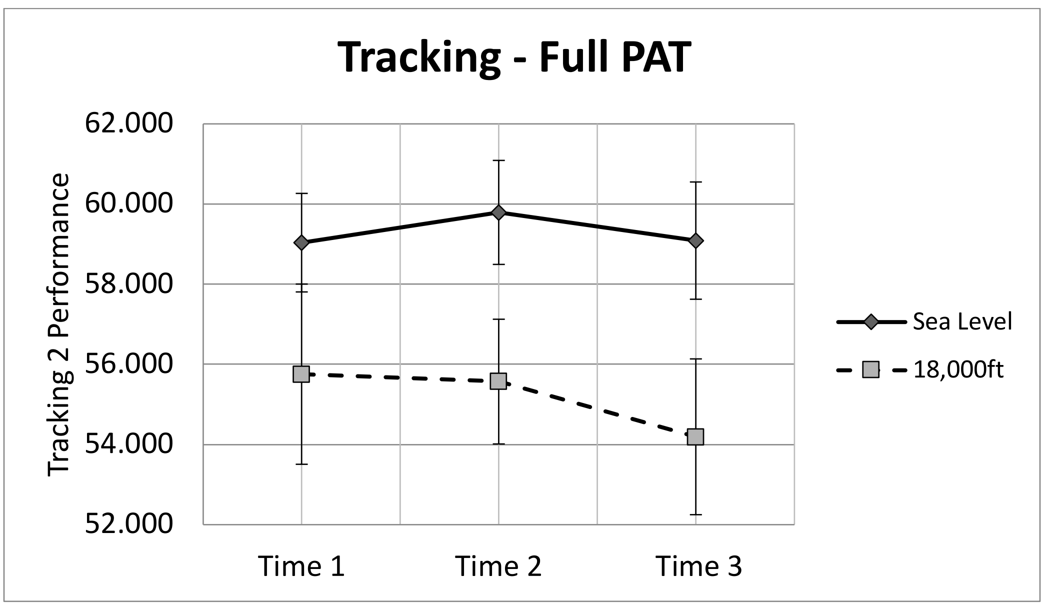

Figure 5: Significant main effect for Stressor $F(2,24) = 15.746$, $p = .001$, $P\eta^2 = .396$. Significant differences between hypoxia and normoxia were found in time points 2 and 3.

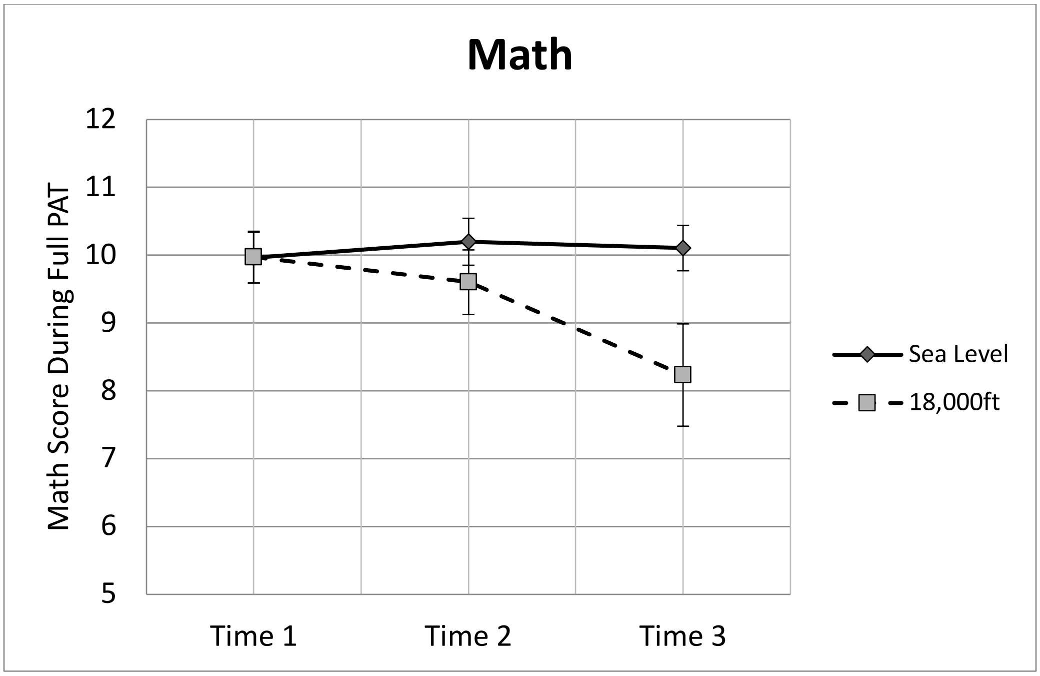

Figure 6: Significant main effect for stressor, F(1,24) = 7.863, p = .010, Pη^2 = .247, significant main effect for time, F(1.597, 36.389) = 3.513, p = .049, Pη^2 = .128, and a significant stressor by time interaction, F(1.465, 48) = 4.465, p = .028, Pη^2 = 15.7. Differences significant at time point 3.

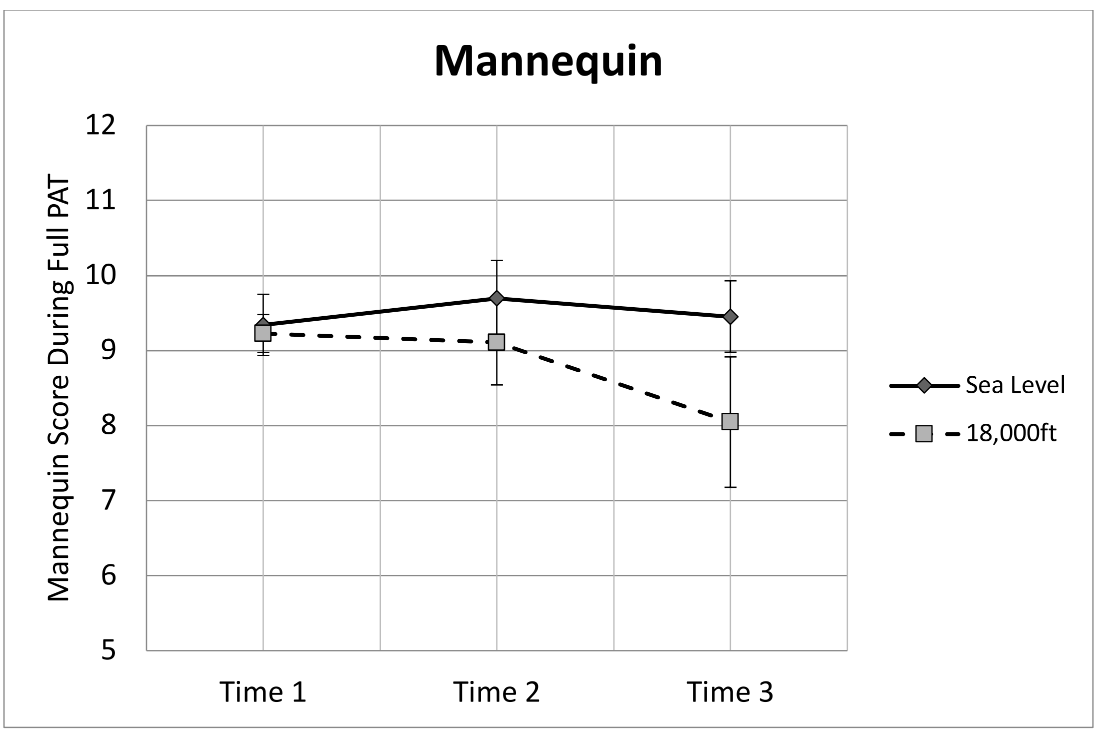

Figure 7: Significant main effect for stressor, F(1.24) = 4.853, p = .037, Pη^2 = .168

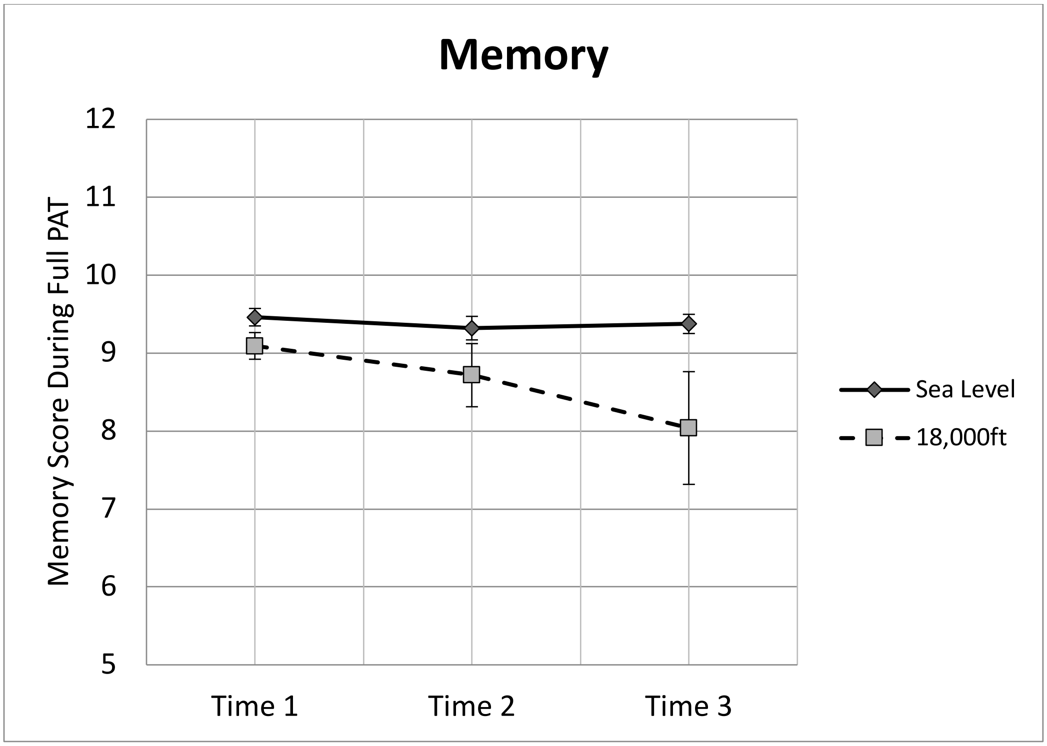

Figure 8: Significant main effect for stressor, $F(1,24) = 4.795$, $p = .039$, $P\eta^2 = 0.167$

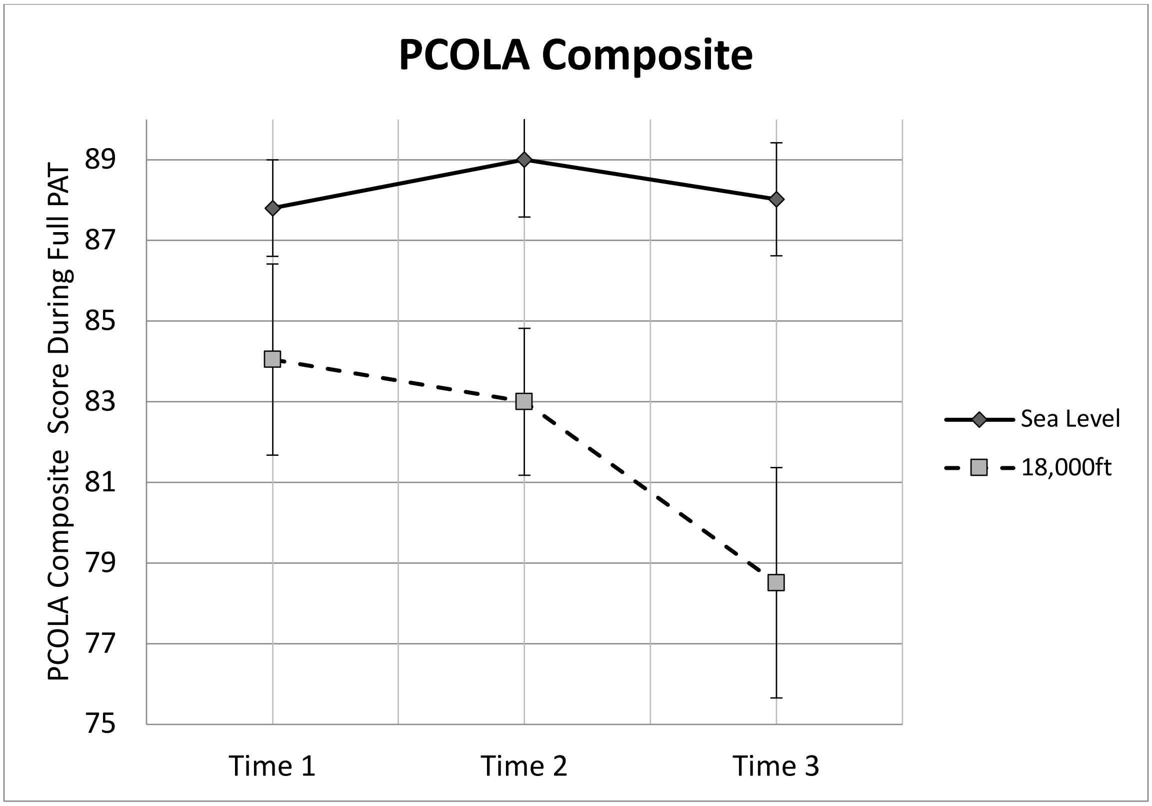

Figure 9: Significant main effect for Stressor, $F(2,24) = 14.630$, $p = .001$, $P\eta^2 = .362$, Significant effect for time, $F(2,48) = 4.373$, $p = .033$, $P\eta^2 = .154$, and a significant stressor by time interaction, $F(1.650, 39.609) = 4.085$, $p = .031$, $P\eta^2 = .145$.. Significant differences between hypoxia and normoxia were found in time points 2 and 3.

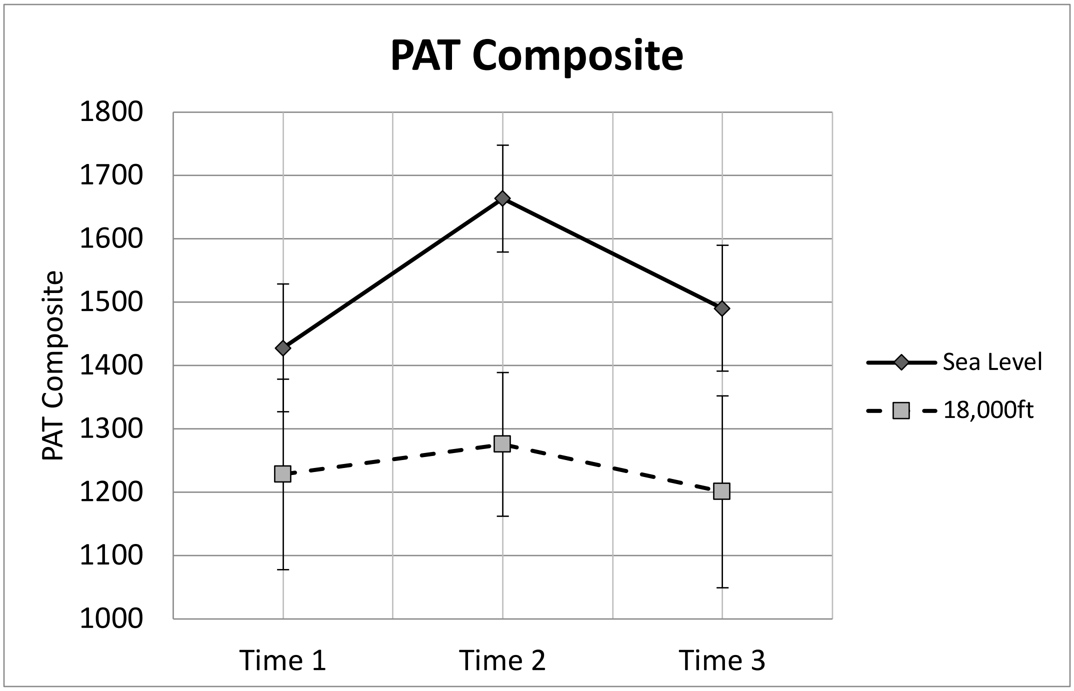

Figure 10: Significant main effect for Stressor, $F(2,24) = 13.627$, $p < .001$, $P\eta^2 = .362$. A Significant main effect was also found observed for Time, $F(2, 48) = 3.477$, $p = .039$, $P\eta^2 = .127$. Significant differences between hypoxia and normoxia were found in time points 2 and 3.

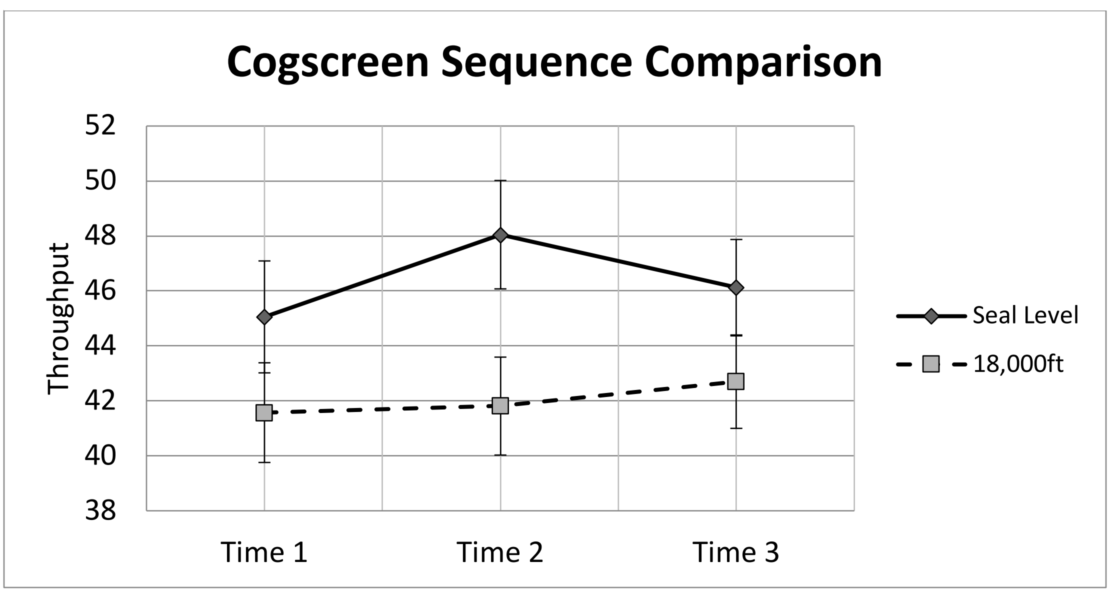

Figure 11: Significant main effect for Stressor, $F(2,24) = 27.218$, $p < .001$, $P\eta^2 = .564$. A significant difference between hypoxia and normoxia was found in time point 2.

Table 5 Summary of effect sizes.

Variable	Effect Size	Stressor Difference at Time Points
CogScreen VSC	0.564	2 Only
Tracking	0.431	2 and 3
PAT Tracking	0.396	2 and 3
PCOLAC	0.379	2 and 3
PAT Composite	0.362	2 Only

Discussion and Conclusion

With regard to our Specific Aim 1, the results of the current study clearly demonstrated that PAT is reliable after adequate training on the tasks with Cronbach's alpha values in the mid-to-upper .80s when measured across metrics on Day 2, which followed the training day (Day 1). While there was variability and relatively low alpha values across Day 1, this suggests that this practice session resulted in asymptotic performance. Importantly, after performance asymptotes, reliability is excellent. Therefore, we would recommend that participants complete around 12 practice sessions before proceeding to experimental testing.

With regard to our Specific Aim 2, the analysis of hypoxia effects across the PAT performance metrics of interest demonstrated significant and large effects for all seven of the variables included. The strongest PAT performance effects were associated with measures of tracking. Tracking only resulted in slightly higher effect sizes than Tracking during Full PAT (see Tables 4 and 5). The observed stressor by time interaction for PCOLAC suggested that hypoxia performance effects, as measured by PCOLAC, get larger as time at altitude increases (see Table 4 and Figure 9). Measures of tracking associated with trials where participants performed the Full PAT also had large effect sizes and significant effects (see Table 4). Specifically, tracking during the Full PAT demonstrated significant differences between hypoxia and normoxia at time points 2 and 3 and importantly large effect sizes (see Figure 5 and Table 5). CogScreen VSC had the largest effect size with significant differences between hypoxia and normoxia detected at time point two only (see Table 4 and Figure 11). Although the CogScreen VSC had the largest effect size, it did not have a significant stressor by time interaction effect, suggesting that it cannot detect the decline in performance associated with hypoxia over time at altitude.

When analyzed independently, the three measures associated with secondary task performance (Math, Manniquen and Memory) each showed significant main effects for stressor, demonstrating that they are each sensitive to the performance effects of hypoxia (see Table 4 and Figures 6-8). Math also had a significant stressor by time interaction. This suggests that Math was also sensitive to the exacerbation of hypoxia effects as time at altitude increased (see Figure 6). Although Mannequin and Memory did not have a significant stressor by time interaction, the data did trend in that direction (see Figures 7 and 8). Given that each session only allows for around 40 seconds of data acquisition for each secondary task, these effects are especially impressive.

The normoxia and hypoxia conditions were not randomized to prevent hypoxia carry-over effects from contaminating performance in the normoxia condition in half of the participants (Phillips et al., 2011). Since normoxia always preceded hypoxia, there is a possibility that cognitive fatigue may explain some portion of the observed effects associated with the hypoxia condition. However, examination of PAT performance across Day 2 does not suggest the presence of significant cognitive fatigue effects to support this alternative explanation. Therefore, it is likely that these observed effects are in fact associated with the hypoxia exposure as opposed to cognitive fatigue.

In summary, this experiment was conducted to determine whether PAT was reliable and sensitive to the performance effects of hypoxia. Data suggested that two PAT performance metrics possessed excellent test-retest reliability (Tracking and PCOLAC) and one metric had acceptable reliability (PATC). This demonstrated reliability shows that these PAT performance metrics should have minimal error variance across experimental trials. Results of the hypoxia experiment show that PAT is sensitive to the performance effects of hypoxia at an 18,000ft normobaric equivalence. All seven of the PAT metrics had a significant hypoxia effect. The PCOLAC and Math variables were also sensitive to the exacerbation of hypoxia performance effects associated with increasing time at altitude. As expected, Cogscreen Visual Sequence Comparison resulted in a high effect size for a main effect of stressor, but unexpectedly, did not show sensitivity to increasing time at altitude, which was shown by PCOLAC and Math. PAT should be considered for use by anyone attempting to detect the performance effects of an aeromedical stressor, such as hypoxia, hypercapnia, hypocapnia, G-stress, hypobaria, fatigue, or motion sickness. There are some changes that will make the PAT an even more powerful performance assessment. The next section presents suggestions for PAT's improvement.

Suggestions to Improve PAT

1. Separate the tracking component from the secondary tasks, so that each component can be performed one task at a time. This will allow participants to get to higher difficulty levels and identify the optimal difficulty for each component for each individual participant over the practice and acquisition phases. Identifying the optimal difficulty level for each individual component within each participant will make effect sizes larger and increase the likelihood that performance effects associated with stressors can be quantified.

2. Consider creating more items for the math task and making these items more challenging. Math items appeared to repeat themselves within sessions and can easily be memorized. Some math items are simply not difficult enough to require participants to carry out mental addition. For example, there are several items that use double digit numbers on the multi- functional display and very small numbers presented on the yellow circles, these items are too easy. Items should be carefully constructed to prevent them from being answered correctly without conducting mental arithmetic. When math difficulty levels increase the presentation pace of the items increases and the allotted time provided for participants to respond decreases as opposed to systematically increasing the difficulty of the items.
3. At times, two sets of memory items were presented following a one-second delay. It is not clear whether this is intended or a bug in the program. It also is not clear whether participants are supposed to respond using both sets of items or just the last set that was presented. Designers should consider using phonemic (i.e., P and T) similarities in item pronunciations to increase difficulty level as opposed to only using larger arrays and faster trial presentations. Developers should also be sure that simple chunking techniques cannot be used to reduce the number of items to be recalled (i.e. ‘B’, ‘A’, ‘D’ becomes one item bad).
4. Present memory items on the MFD as opposed to a red distractor. This will provide consistent areas of interest for math, mannequin and memory to afford for computing better eye-tracking metrics.

References

Ponterotto, J.G. & Ruckdeschel, D.E (2007) An overview of coefficient alpha and a reliability for estimating adequacy of internal consistency coefficients with psychological research measures. *Perceptual Motor Skills*, 105, 10023-7478.

Cohen J. (1988). *Statistical Power Analysis for the Behavioral Sciences*. New York, NY: Routledge Academic

Kerlinger, F. N. (1973). *Foundations of behavioral research* (2nd ed.). New York: Holt, Rinehart and Winston. Loesch, L. C., & Vacc, N. A. (Eds.) (1997). Research in counseling & therapy. Greensboro, NC: ERIC Counseling and Student Services Clearinghouse.

Rice GM, Moore JL, Moore JL Jr., Kay G. (2003). Validation of a cognitive test battery for assessing hypoxia related cognitive dysfunction. *Aviation, Space and Environmental Medicine*, 74, 281.

Nevo, D. (1985) Face Validity Revisited, *Journal of Educational Measurement*, 22, 287-293.