From Lab to Battlefield: Exploring the Relationship Between Military and Basic Science Tasks for Measuring Competencies

William Stalker, Summer Rebensky, Ramisha Knight, Samantha Perry, Shawn Turk,
Aptima, Inc.
Fairborn, OH
lstalker@aptima.com, srebensky@aptima.com, rknight@aptima.com, sperry@aptima.com, sturk@aptima.com

Quintin Oliver, Winston "Wink" Bennett
Air Force Research Laboratory
WPAFB, OH
quintin.oliver@us.af.mil, winston.bennett@us.af.mil

ABSTRACT

Many scientists have struggled to walk the line between designing an experiment with rigorous control and conducting research whose findings widely generalize. As government and industry invests funding into developing training systems and competency measurement approaches, the importance of transferability of control study findings will be key to prevent the valley of death. This debate often shows up when practitioners are challenged to translate experimental tasks from basic research to a more applied context. The *n*-back (Kirchner, 1958), a respected measure of working memory ability, is commonly poked fun at for disconnect from any realistic task. Recent work (e.g., Vine et al., 2020 as cited in Vine et al., 2021) has furthered these concerns. The present study expands on an adaptive training system approach presented in 2022 (Rebensky, et al., 2022), which explores the *n*-back's transferability via a domain relevant task, dubbed Mission Relevant Audio Cue Task (MRACT), which has been designed to mimic communication call-and-response procedures of military personnel that would challenge working memory similar to the *n*-back. Participants were recruited and completed both tasks separately in the Driving Adaptive Research Testbed (DART) developed at the Air Force Research Laboratory. The data from these tasks were compared along relative difficulty of task, perceived difficulty (NASA-TLX), and their demands on physiological measures related to mental workload (fNIRS and heart rate variability). The findings of this study reveal similarities between the two tasks but also points of substantial divergence, which have implications for basic research laboratories that train algorithms within adaptive training settings. These conclusions lead to suggestions for future practitioners looking to walk the same line when developing domain relevant tasks based on basic research. The paper and presentation will provide guidance for research and development labs on the design of tasks to test novel measurement and algorithm approaches.

ABOUT THE AUTHORS

William (Liam) Stalker M.S., Aptima, Inc., is an Associate Scientist in the GRILL®. He uses his proficiency in neuroergonomics, human factors, and experimental psychology to the aid the Training, Learning, and Readiness Division. His efforts focus on adaptive training in virtual testbeds as well as survey methodology and implementation. William is a doctorate candidate and has earned his MS in human factors / industrial and organizational psychology from Wright State University and his BS in neuroscience from the Ohio State University. William's current graduate studies focuses on (1) understanding how one's visual environment impacts performance, and (2) how adaptive learning systems can be paired with physiological measures to improve outcomes.

Dr. Summer Rebensky, Aptima, Inc., is a scientist who has a background focusing on human performance, cognition, and training in emerging systems. Dr. Rebensky has previous experience as a research fellow as a part of the Air Force Research Laboratory conducting research on drone operations and human-agent teaming utilizing game-based technology. Her research experience involves leveraging Virtual Reality (VR), Augmented Reality (AR), and game-based technology to optimize human performance in training and operations. Dr. Rebensky received her BA in psychology, MS in aviation human factors, and PhD in aviation sciences focused on human factors from Florida Tech.

Dr. Ramisha Knight, Aptima, Inc. is a scientist who specializes in human cognitive neuroscience and experimental psychology. She has experience using a broad range of neurophysiological techniques to measure cognitive processes and the neural architecture associated with visual attention and perception. Her research includes statistical methods and machine learning-based approaches to predict behavior and performance. Prior to Aptima, Dr. Knight completed her postdoctoral work at the University of Illinois at Urbana-Champaign and the Beckman Institute for Advanced Science and Technology. She holds a PhD in psychological and psychiatric science from Università degli studi di Verona (Italy), an MS in cognitive neuroscience from the University of Durham (England), and a BA in psychology from Hawaii Pacific University. *US Citizen.*

Dr. Samantha (Baard) Perry, Aptima Inc., is a senior scientist and the deputy director of the Training, Learning and Readiness Division, which focuses on understanding, capturing, and assessing individual and team processes, states, and performance. She has more than 15 years of academic and applied research experience with the Air Force, Army, and NASA with expertise in adaptation, motivation, training design and evaluation, and unobtrusive measurement of individual and team processes, states, and performance. Dr. Perry holds a PhD and MA in industrial and organizational psychology from Michigan State University and a BA in psychology from George Mason University.

Shawn Turk, Aptima Inc., is an associate software engineer in the GRILL®. and leverages a BS in digital simulation and game engineering technology from Shawnee State University to explore training solutions utilizing artificial reality and virtual reality technology. He has utilized Unreal Engine and Unity to incorporate sensor data to augment and adapt virtual tasking.

Quintin Oliver, AFRL, is a Computer Scientist in the GRILL®. His work utilizes virtual, augmented, and mixed reality technologies to create rapid prototypes of environments focused on personalized training. In these environments, he leverages his interests of 3D modeling and artificial intelligence to create unique experiences.

Dr. Winston "Wink" Bennett, is a Senior Principal Research Psychologist and Readiness Product Line Lead for the Warfighter Readiness Research Division, Airman Systems Directorate, 711th Human Performance Wing, Air Force Research Laboratory. Wink is a recognized leader in education, training, competency definition and assessment, and performance measurement research. He has been involved in a number of multinational research collaborations and continues to support collaborations around the world. He is a Fellow of three distinguished professional societies and the Air Force Research Laboratory.

From Lab to Battlefield: Exploring the Relationship Between Military and Basic Science Tasks for Measuring Competencies

William Stalker, Summer Rebensky, Ramisha Knight, Samantha Perry, Shawn Turk,
Aptima, Inc.
Fairborn, OH
lstalker@aptima.com, srebensky@aptima.com, rknight@aptima.com, sperry@aptima.com, sturk@aptima.com,

Quintin Oliver, Winston "Wink" Bennett

Air Force Research Laboratory
WPAFB, OH
quintin.oliver@us.af.mil, winston.bennett@us.af.mil

INTRODUCTION

The beginning of one's journey between the academic and applied world starts with the realization that it is difficult to translate the findings of "basic" laboratory research to one's desired context. Many of the carefully devised cognitive tasks explored in academia are conducted in an unrealistically barren or carefully controlled environment, lack the multi-tasked nature common in most operational settings (Vine et al., 2021), and are tested disproportionally with Western, Educated, Industrialized, Rich, and Democratic (WEIRD) people (Henrich et al., 2010). These hurdles tend to cause practitioners frustration, which then eventually builds up into an inevitable opinion piece spouting that "nothing applies" and we "need new tasks." While we support the development of new cognitive tasks, this is not meant to be harsh criticism of the traditional toolkit of academic literature. One might argue that it makes more sense for the scientific community to build upon the stripped-down versions of cognitive tasks rather than trying to twist an applied version of an experimental task to work for a different operation. Twisting these tasks without a full understanding of the fundamentals and the theory behind them is likely part of the reason why some results appear to not transfer. For example, one should be careful to not misuse scanning behavior, an indicator of cognitive workload in pilots (Tole et al., 1982), as a measure of workload in a task where there is no reason to expect that deviation from a set scanning would indicate a change in workload. Measures of a task can translate to a new task when they are highly similar to each other, such as the scanning behaviors of trained pilots and surgeons (e.g., Lim et al., 2023), but transfer is still not guaranteed. The present study explores the transferability of a basic task, the *n*-back (Kirchner, 1958), to a domain relevant task dubbed Mission Relevant Audio Cue Task (MRACT). This work expands on an adaptive training system approach presented in 2022 (Rebensky, et al., 2022) and further explores the relationship from military and basic science tasks for measuring competencies. From the lab to the battlefield, comparison of the two tasks reveal similarities but also points of substantial divergence which have implications for basic research laboratories that train algorithms within adaptive training settings.

METHODS

Participants

A total of 28 participants were recruited from the Dayton Metropolitan Area. All participants were 18 years or older, and U.S. citizens with normal or corrected hearing. Participants were compensated $40 for participating. Demographic information, experience with driving, their hobbies, and whether they had recently consumed any stimulants such as nicotine or caffeine was collected following consent to participate in the study and prior to beginning the study. The overall median age was 25.5 years old (18 – 68). The majority of the participants drove on average more than 5 hours per week (65%) and over half of the participants experienced little to no anxiety driving (70%). In addition, most of the participants regularly played video games (72%).

Experimental Testbed

The data was collected from the dual task experimental testbed, the Driving-based Adaptive Research Testbed (DART). A project first theorized in Rebensky et al. (2022), the DART was developed at and is hosted by the Gaming Research Integration for Learning Laboratory (GRILL®; See Figure 1). The DART acts similarly to traditional

adaptive systems; it adapts the difficulty of the task based on the user's state and their previous performance. The DART was initially designed with two goals in mind. First, an engineering goal intended to demonstrate that Unreal Engine can receive live-streamed physiological data related to mental workload and that it can dynamically update the environment and tasks based on the user state. Second, a scientifical goal to determine what physiological responses are most effective to adapt a simulation environment based on performer state (the results of goal 2 can be found in Stalker et al., 2024). This testbed dually challenges its users, requiring them to drive through a simulated hostile terrain using a mock-HMMWV while simultaneously engaging in various cue discrimination tasks. The initial instantiation of DART uses multiple measures of mental workload, including a modified NASA-TLX (Hart & Staveland, 1988), a functional near-infrared (fNIRS) sensor developed by BionicaLabs called NIRSense, and a Polar H10 Heart Rate Monitor sensor for tracking blood flow and heart rate (HR), and performance metrics from the simulated tasks themselves.

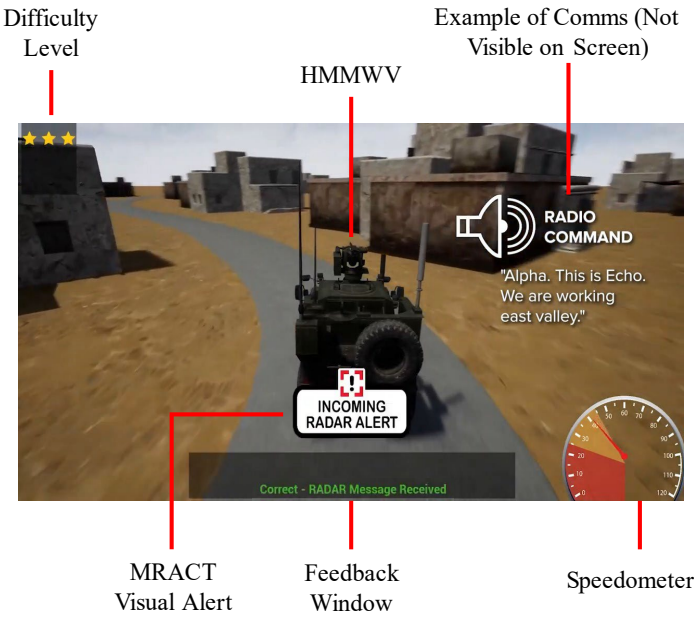

Figure 1. DART Testbed & Elements

Experimental Design and Procedure

The experiment consisted of a short Qualtrics survey (5 minutes) followed by training and a longer behavioral task (45 minutes) that consisted of three phases. Participants completed all steps on site. After consenting to participate, subjects were outfitted with a fNIRS sensor developed by BionicaLabs called NIRSense, and a Polar H10 Heart Rate Monitor sensor for tracking blood flow and heart rate. Subjects were then asked to complete a short demographic Qualtrics survey. After completing the Qualtrics survey, participants reviewed an instructional slide deck for the behavioral task for controlling the vehicle. Participants were then informed that their driving performance score during the experimental task would be based on how well they managed to stay on the road and the speed they maintained while doing so. Participants were reassured that they would be returned to the road automatically should they drive off after 5 seconds or they could also choose to reset themselves at any given time by selecting a specific button. Participants were then asked to drive along a 3 minute straight road driving baseline for physiological metrics, and then given more instructions for the secondary cognitive task that would be performed while driving.

Task 1: Basic Research Task – Auditory *n*-Back

The first cognitive task was a form of the classic *n*-back task (Kirchner, 1958). As detailed in past work (Stalker et al., 2024) and briefly described here, participants in this case were required to perform a 2-back auditory memory task. Letters were played aloud through the participant's headset every 2.5 seconds in a random order. Participants were instructed to press either of the paddles on the steering wheel when a letter matched the same letter played 2 letters back. Participants completed this task while driving along a series of different levels of road curviness. Participants first completed one 2-minute segment with 46 *n*-back letters played with the first trial on a "medium" level difficulty or somewhat curvy road. Then, participants completed five additional 2-minute trials (12-minutes total). After each trial, the adaptation algorithm would run and determine whether to make the roads easier (straighter) or harder (more curved) to drive. Before beginning the next trial, participants would fill out mental workload prompts from the

modified NASA-TLX (Hart & Staveland, 1988). After completing the basic research task, participants were introduced to more instructions for the applied mission task.

Task 2: Applied Mission Task – Mission Relevant Auditory Cues Task (MRACT)

Participants were introduced to a new cognitive task that was designed as a part of this study to be more generalizable to a military driving mission context. Operational military settings cognitive demands are complex, oversaturated in signals, and require understanding relevant cues and responding appropriately. As demonstrated in previous research, military cognitive tasks do not often correlate closely to basic tasks such as the *n*-back (Vine et al., 2020). The *n*-back, while difficult, requires keeping a running list in working memory, which is an uncommon cognitive demand in military settings. Therefore, we sought to design a more applicable testing scenario by building upon previous more operationally relevant cueing tasks similar to those mentioned in Eddy et al. (2015), Lenné et al. (2014), and the Military Specific Auditory *n*-Back task (MSANT; Vine et al. 2020). The MSANT, for example, requires users to press specific inputs on a keyboard or game-based steering wheel when presented with different operational relevant sounds and withhold input during non-relevant cues and when following radio comm procedures (e.g., specific gunfire or radio messages). These previous instantiations provided guidance for the development of the MRACT.

The MRACT provided a mix of (a) auditory tones (beeps) that signaled that an immediate input was needed, and (b) radio communications that explicitly asked for an input, sometimes immediately or sometimes after a brief delay. Regarding the auditory tones, two distinct audio cues required responses, the IED Warning Cue and the Radar Cue. The IED Warning Cue was a series of seven rapid short beeps played over two seconds that indicated that the participant needed to press the left paddle on the steering wheel as soon as possible. The Radar Cue was a rapid series of two low-toned beeps played over two seconds indicating to the participant that they needed to press the right paddle on the steering wheel as soon as possible. Extraneous tones that were not either of these distinct cues were to be ignored with no input given. Regarding radio communication, participants ignored the callout until their callsign was spoken. When spoken to, the participant acknowledged the callout by pressing the "A" button. The radio communication message either required the participants to immediately press a button on the steering wheel in response, or briefly delay the input until heralded again (see example breakdown of require inputs in Table 1). Similarly to the *n*-back portion of the experiment, the MRACT required participants to drive along a series of roads.

Table 1. MRACT Cues & Appropriate Responses

Audio Type	Message	Response
Radar Tone	*(low slow beeped tone)*	Press right paddle
IED Tone	*(rapid high pitch tone)*	Press left paddle
Extraneous Tones	*(All other tones)*	No response
Comms Message (Acknowledge only)	*"Bravo this is Alpha, I am entering building 119"*	Press "A" to acknowledge
Comms Message (Acknowledge & Respond)	*"Bravo this is Delta, press X to disable drones"*	"A & X"
Comms Message (Acknowledge & Delayed Response)	*"Bravo this is Charlie, when I'm ready, press Y"* *…* *"Bravo I am Ready"*	"A," …. "A & Y" later when signaled
Comms Message (To anyone other than Bravo)	*"Delta this is Alpha, I am entering building 119"*	No response

Participants first practiced on medium-difficulty roads (slight curves) while completing only the auditory tones portion of the MRACT in order to incrementally introduce them to the tasks of the MRACT. Participants completed one 2-minute practice segment where one cue played every 2-4 seconds. Random cues that did not require responses as well as radio comms played alongside the distinct cues that the participants needed to respond to. Next, participants were introduced to the radio communications portion of the MRACT . Participants responded to comms messages to their designated callsign (Bravo) by pressing the "A" button on the steering wheel to acknowledge all radio calls. Some radio calls asked for another button input, asking the user to press either the "X," "Y," or "B" button after they had acknowledged the caller using the "A" button. Further, to simulate the *n*-back style of tasking, some radio calls asked participants to input the "X," "Y," or "B" button once the person on the radio comms was ready—simulating a delayed responses style message. Users were instructed to (a) acknowledge the first radio call with the "A" button, (b) remember the specific "X", "Y", or "B" button input that would be required, (c) wait to hear the matching radio call from the same individual instructing the participant they were ready for the command, (d) press "A" to acknowledge again, and (e) finally press the desired command established in the first radio call (see example breakdown of required inputs in Table 1). Radar/IED tones and comms messages from other individuals could play between the two delayed response messages.

Participants were then tasked with practicing driving along a series of hard-difficulty roads (intense curves) while completing the MRACT. Adaptations also occurred during the MRACT, but in this task, adaptations were the presence or absence of multi-modal cueing for the auditory tone portion of the task (no additional cues, static visual cueing, flashing visual cueing; starting with static visual cueing). Participants were instructed to complete five more 2-minute trials (12-minutes total). After each trial, the adaptation rule-based logic would determine whether to add redundant visual warning cueing to the auditory warning task or to remove it. The rationale for multimodal cueing was two-fold: (1) in a training environment visual cueing could provide scaffolding and be slowly removed as trainees are better able to handle the operational mission, and (2) could be implemented within augmented reality or head mounted displays in an operational vehicle and therefore were reasonable modifications to the environment that could take place in a mission setting to combat high workload effects.

After completing the two tasks, participants finished the experiment by driving along a simulated straight road for 3 minutes to collect a final baseline from the physiological sensors. Participants then removed the sensors and were debriefed. The entire process took approximately 60 minutes. A procedure flow can be seen in Table 2.

Table 2. Study Procedure

Stage	Task	Description
Setup	Study Prep	Informed consent, sensor fitting, demographic survey
	Training & Task Instructions	Text on sim screen & slides shown in simulator
Baseline	Baseline	Straight road + TLX prompt
N-back	2-back Practice	(Letter played every 2.5 seconds) + TLX prompt
	2-back w/ Road Adaptation (x5)	(Letter played every 2.5 seconds) + TLX prompt
Mission	Tone Practice	(Sound played every 2-4 seconds) + TLX prompt
	Tone + Comms Practice	(Sound/audio played every 2-4 seconds) + TLX prompt
	Tone + Comms	(Sound/audio played every 2-4 seconds) + TLX prompt
	Tone + Comms w/ multimodal adaptation (x5)	(Sound/audio played every 2-4 seconds) + TLX prompt
Post	Post baseline & end study	Collect post baseline and end study

Measures of Performance and State

Measures of participant performance included: (a) The average miles per hour driven, (b) the number of times even one wheel went off road, (c) the number of manual or automatic (5 seconds of the vehicle fully off road) resets on the road, (d) the percentage of correct responses for both the n-back task and the MRACT task. Metrics of participant

state included: (a) fNIRS oxygenation levels, (b) Polar H10 Heart Rate, (c) One-question mental workload prompt from the modified NASA-TLX. Each of these served as a measure of mental workload. More details of the specific measures and how they adapted the simulation environment are described in more detail in Stalker et al. (2024). Participants shifted through easy, medium, or hard conditions dependent on performance and state. In the easy condition, participants received straighter roads in the n-back task and flashing multimodal assistance in the MRACT task. In the medium condition, participants received slightly curvy roads in the n-back task and static multimodal assistance in the MRACT task. In the hard condition, participants received curvy roads in the n-back task and no additional visual assistance in the MRACT task. In all conditions for MRACT, participants drove on medium roads.

RESULTS

Descriptive statistics can be seen in Table 3 by task and by difficulty level for each of the measures captured. Multiple two-way multivariate ANOVAs (MANOVA) were conducted to determine the impact of Task type (*n*-back vs MRACT) and level of difficulty (Easy, Medium, Hard) across task performance measures (Average MPH, Times Off Road, Resets, and Accuracy) and physiological measures (fNIRS HbO2, fNIRS HbD, Heart Rate) and the self-report workload measure (NASA-TLX). Means and standard deviations were calculated and are shown in Table 3. Multiple multivariate test statistics were calculated to test the statistical significance of the different effects of the two independent variables, Task and Difficulty, shown in Table 4. First, the multivariate result was statistically significant for Task ($p < 0.001$). Second, the multivariate result was statistically significant for Difficulty ($p < 0.001$). As seen in *Task x Difficulty*, there is a statistically significant interaction effect between task and difficulty, $F(16, 427) = 3.0120$, $p < 0.001$; Wilks' $\Lambda = 0.834$.

Table 3. Descriptive Analysis Results

	n-back						*MRACT*					
	Easy		*Medium*		*Hard*		*Easy*		*Medium*		*Hard*	
	M	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*	*M*	*SD*
Average MPH	67.1	13.8	60.2	15.6	52.8	11.9	60.9	11.6	61.0	9.7	56.5	11.1
Times Off Road	18.3	8.5	14.3	9.2	11.5	8.3	20.7	7.4	18.7	8.8	12.0	0.2
Road Resets	1.0	1.4	0.4	0.8	0.2	0.6	0.4	0.8	0.2	0.4	0.2	0.6
% Correct	88.3	12.3	89.6	10.4	90.4	6.0	80.7	17.9	85.6	16.5	83.4	18.6
Polar HR	76.27	12.19	77.15	14.14	80.22	13.01	78.52	13.97	76.51	12.93	77.53	11.56
fNIRS HBD	-7.0E-05	2.0E-04	8.6E-07	2.4E-04	1.3E-05	1.8E-04	-4.8E-05	1.7E-04	6.8E-05	3.0E-04	3.8E-05	2.1E-04
fNIRS HBO2	4.7E-04	8.2E-04	6.7E-05	8.1E-04	-6.2E-05	5.6E-04	5.1E-04	7.9E-04	-5.3E-05	1.2E-03	-1.1E-04	7.4E-04
TLX	12.2	4.5	12.4	3.5	12.5	2.8	13.1	3.9	13.7	3.5	13.6	3.0

Table 4. Multivariate Analysis Results

	Task				*Difficulty*				*Task x Difficulty*			
	Value	*F*	*df*	*p*	*Value*	*F*	*df*	*p*	*Value*	*F*	*df*	*p*
Wilks' lambda	0.8009	8.1711	8, 263	< 0.001	0.8311	3.1869	16, 526	< 0.001	0.8341	3.1202	16, 526	< 0.001
Pillai's trace	0.1991	8.1711	8, 263	< 0.001	0.1732	3.1289	16, 528	< 0.001	0.1676	3.0178	16, 528	< 0.001
Hotelling's trace	0.2486	8.1711	8, 263	< 0.001	0.1981	3.2473	16, 427	< 0.001	0.1968	3.2248	16, 427	< 0.001
Roy's Largest Root	0.2486	8.1711	8, 263	< 0.001	0.1674	5.5250	8, 264	< 0.001	0.1856	6.1264	8, 264	< 0.001

Follow-up univariate analyses showed both fNIR HBD and HBO2 were statistically significantly impacted by Difficulty ($p = 0.007$, $p < 0.001$, respectively) with no significant difference for task nor an interaction effect. Regarding task performance measures, Average MPH and Times Off Road were also statistically significantly related to Difficulty ($p < 0.001$, $p < 0.001$, respectively). Multiple variables were significantly different related to task, meaning the variables differed between *n*-back and MRACT, including NASA TLX ($p = 0.0130$), Percent Correct ($p = 0.0003$), Times Off Road ($p = 0.0498$), and number of Road Resets ($p = 0.0057$). Number of Resets ($p = 0.0197$) and the Average MPH ($p = 0.0165$) both showed interaction effects between Task and Difficulty level. Univariate analyses are presented in Table 5, with the implications of these findings discussed in the following section.

Table 5. Univariate Analysis Results

	Task				*Difficulty*				*Task x Difficulty*			
	SS	*F*	*df*	*p*	*SS*	*F*	*df*	*p*	*SS*	*F*	*df*	*p*
Average MPH	7.4613	0.0484	1	0.8260	4212.4941	13.6633	2	0.0000	1285.6519	4.1700	2	0.0165
Times Off Road	302.5370	3.8827	1	0.0498	3175.0106	20.3739	2	0.0000	163.3018	1.0479	2	0.3521
Road Resets	5.1678	7.7677	1	0.0057	15.1821	11.4101	2	0.0000	5.3045	3.9866	2	0.0197
Percent Correct	2822.9494	13.5574	1	0.0003	428.9583	1.0301	2	0.3584	135.1183	0.3245	2	0.7232
Polar HR	19.2741	0.1150	1	0.7347	181.8126	0.5426	2	0.5819	308.4859	0.9206	2	0.3395
fNIRS HBD	7.8450e-08	1.7362	1	0.1887	4.6225e-07	5.1151	2	0.0066	2.3418e-0-8	0.2603	2	0.7711
fNIRS HBO2	9.4508e-08	0.1518	1	0.6971	1.8467e-05	14.8289	2	0.0000	2.497e-07	0.2004	2	0.8185
TLX	78.6237	6.2534	1	0.0130	8.8718	0.3528	2	0.7030	0.6314	0.0251	2	0.9752

DISCUSSION

These discrepancies can lead to difficulties for practitioners who experience different results when existing academic research does not directly apply to their contexts, sometimes resulting in calls for new tasks. While we recognize the importance of developing novel cognitive tasks, it is essential not to dismiss the value of established ones. Instead, a more productive approach would be to build upon the fundamental principles and theories underpinning these tasks before attempting to adapt them to new contexts (Rebensky et al., 2022). Analyses showed that the measures selected for comparison were significantly impacted by the task type, the difficulty level of the task, and the interaction between these two factors. The interpretations of these findings and more are expanded upon in the sections below.

Generalizability: Measures Robust to Changing Tasks

Researchers and practitioners alike are often tasked with finding measures that prove robust across multiple settings. A subset of this goal was to discover measures that are impacted by difficulty but not by task. The fNIRS data and two of the selected task performance measures (Average MPH & Times off Road) met this criterion. The findings from this research suggest that fNIRS may be an ideal candidate that transits from the laboratory to operational settings. Both fNIRS readings of HBD and HBO2 were statistically significantly impacted by task difficulty but not the type of task. The significance only on changing difficulty implies that fNIRS measures were not impacted by the transition between a traditional basic experimental task and a mock applied task, but only by the increase in difficulty. This suggests that fNIRS measures may be suitable when looking to translate findings between basic and applied settings. Regarding task performance measures, Average MPH and Times Off Road were also statistically significantly related to difficulty but not type of task. Participants drove slower and off the road more often in the higher difficulty conditions, regardless of task. This was likely due to a component of the adaption logic made the overall course curvier in higher difficulty trials, limiting the participants' ability to drive quickly if they were to maintain high proficiency.

Limitations: Measures More Impacted by Task & Design

Some of the investigated measures were statistically significantly impacted by the type of task. These included ratings of workload collected from a modified NASA TLX, and task performance measures such as Percent Correct, Times Off Road, and number of Road Resets. Self-reported workload via the modified NASA TLX was significantly higher

in the MRACT than the traditional *n*-back task. The precise reason for this is unknown, but one may speculate that this is likely partially a result of the MRACT having more components to it than the *n*-back task. People are known to rate tasks with more components as more difficult than tasks with less, even when performance data suggests otherwise (Vidulich & Tsang, 2012). Although they were designed to exert similar cognitive demands, the MRACT required participants to remember three different types of inputs while the *n*-back task only required one. Alternatively, or potentially additionally, the story aspect of the applied task may have led participants to be more invested and therefore led to them committing more resources to the task, leading to high ratings of perceived workload.

A few task performance measures correlated with the type of task. While statistically different, the difference in Percent Correct between the two tasks is likely due to task differences. Different tasks are likely going to vary in the Percent Correct. For example, field goals in American football and penalty kicks in soccer are somewhat comparable tasks but ultimately statistically differ in success rate. Considering different tasks have different success rates and implications for any errors (e.g., a wrong response on an n-back has no repercussions, but a missed fire in a simulated air-to-air engagement does) one would anticipate not only different motivation levels and performance, but also differences in scores.

Times Off Road was also comparatively higher in the MRACT condition. It is hypothesized that the MRACT condition required participants to make a greater variety of inputs than the *n*-back task condition. In the n-back condition, either the left or right paddle was a sufficient input. However in the MRACT condition, participants had to press specific paddles as well as steering wheel inputs, which when turning, may have been difficult to do so. Participants may have prioritized using the controller inputs and therefore more hesitant to turn the wheel which would disrupt their ability to successfully respond to the audio cues. If this is true, a takeaway from this finding would be a reminder to try to keep the input modalities the same across tasks one wishes to compare. In future iterations, we anticipate updating the tasks to verbal inputs due to improvements in speech recognition technologies. The last performance measure to correlate with the type of task was the number of Road Resets which was greater in the *n*-back task condition. A potential reason for this could be the difference is road curviness changes. In the *n*-back task condition, the course became more or less curvy. In the MRACT condition, the curviness of the roads did not change; the amount of multimodal cuing changed between the difficulty levels instead. Changing the amount of multimodal cuing was included in interest of exploring workload-based augmentation. This however may have led to the unintended difference between the conditions seen here.

Number of Resets and the Average MPH both showed interaction effects between task type and difficulty level. Number of Resets was higher in the Easy level of difficulty for both task conditions. This is likely a floor effect. Easy might have been too difficult for some participants leading them to still perform poorly, resulting in a high number of resets, despite being on the lowest difficulty level. The floor effect could be alleviated by including more difficulty options and will be explored in future data collections. Average MPH showed little variance in the MRACT condition but did vary substantially in the *n*-back task condition based on difficulty. Average MPH was highest in the *n*-back task easy condition and lowest in the hard condition. This difference likely can be attributed to the difference in curviness of the road. The hard condition in the *n*-back had much curvier roads, putting pressure on the participant to reduce their speed in order to stay on the road. The lack of variance in the MRACT is likely because the curviness of the road did not adapt.

While a subset of the research goal was to discover measures that are impacted by difficulty but not by task, there were bound to be some measures that failed to meet this criterion. Heart rate was not statistically significantly impacted by either independent variable or their interaction. This finding conflicts with past research detailing linked relationship between the construct of mental workload and various heart related measures (Mehler et al., 2009). This leads us to believe that heart rate is likely sensitive to these independent variables but is too diluted from noise caused by other factors. Heart rate can be impacted by many different things including factors external to the experiment like outside clamor and prior stimulant use, and internal factors such as perceived stress and the shaking of the steering wheel. These factors and the fact that heart rate readings were averaged across the two-minute trial likely contributed to its diagnostic inabilities.

Despite collecting data from participants from a broad range of background experience, very few participants stayed consistently in the Medium level of difficulty. Many of the participants (n = 14) finished in the Easy or Hard level. Based on the results that many participants ended up in the easy or hard difficulties and the challenges discussed above, more variance in difficulty will be incorporated into future data collections. The team plans to explore

generative methods to create more variance levels. Another path that will be explored is that the weighting of factors in the adaptation logic should be revised. Based upon adaptive system design recommendations in (Rebensky, 2022), we plan to pursue more advanced adaptation logic via artificial intelligence now that data is available. Researchers and designers of adaptive systems should consider adding more difficulty tiers, often in both directions, to avoid ceiling and floor effects.

Recommendations & Future Directions

Although many relationships that were significant were a result of the design of the testbed, it still brings forth a valuable point—tasks that are demanding are not the same as operational missions. Considering the development of new technologies to broadly support Department of Defense (DoD) training, ones that are robust to services and career fields will provide the greatest value. Therefore, our recommendations and future directions are as follows:

Exploring more generalizable constructs

Having the same performance measures reduces the likelihood that the tasks are tapping into diverging skillsets and offers opportunities for easier comparison. One approach would be to find performance measures that match one another in the test environment and operational environment. If the same performance measures cannot be used, one should look to more generalizable performance measures that can be contextualized in multiple missions (e.g., error rate, reaction time). fNIRS measures and some measures of the environment showed promise as generalizable measures. Although more research is needed to confirm these findings, it provides support for moving towards a more construct level design. Some measures like measures of performance, might have wide differences between them. However, other constructs such as accuracy might be more extensible to other task environments. (e.g., number of times someone successfully shoots a goal and land an aircraft are very different scenarios, but distance from the center for both scenarios may be an extensible measure). Future research also aims to explore other physiological metrics. One being more temporally sensitive heart related measures and seek to isolate the impact of extraneous factors on heart rate related information.

Designing operationally relevant tasks and adaptations

The initial design of DART intentionally created differences between the two environments. The adaptations (curvy roads) in the n-back condition were based upon the most common adaptations administered in adaptive driving research (Zahabi et al., 2020). Whereas the design of MRACT adaptations aimed to lean into adaptations relevant in the operational environment which could take one of two forms: (1) adapting to increase the difficulty by creating new challenges that simulate military relevant sudden changes (e.g., new enemy spotted), or (2) adapting to offload operator workload through multimodal design (Giang et al., 2010). Considering the growth of artificial intelligence over the past couple of years and the potential for artificial intelligence to support the warfighter in intense moments, the team opted for style 2. The different adaptation styles resulted in many of the differences observed. The findings still shed light as to when algorithmic systems trained on controlled settings may or may not be able to transfer to operational environments—which highlights the importance of maintaining consistency if one desires near perfect transfer between types of tasks, even when the change seems subtle. However, future data collections may benefit from maintaining the same adaptive logic for each task type to provide insights if there are unclassified, basic style tasks with similar adaptations that can extended to the operational environment. In future research, we plan to explore adapting environments by incorporating realistic adaptations to operational missions (e.g., obstacles moving into the road) in a style more in line with approach one. For other systems, researchers should be cautious to avoid introducing adaptation factors that reduce the ability to map operational measures. Measures that fluctuate with the scenario adaptations are more likely to become covariates, such as road curviness's impact on average MPH and times off road observed here.

CONCLUSION

The transition from academic research to practical applications can be challenging due to the noticeable discrepancies between controlled laboratory experiments and real-world operational settings. Academic studies often investigate cognitive tasks in simplified environments, lacking the complexity and multi-tasking demands commonly encountered in various domains (Vine et al., 2021). Additionally, these investigations primarily involve Western, Educated, Industrialized, Rich, and Democratic (WEIRD) populations (Henrich et al., 2010), which may limit the generalizability of findings. The present study examines the applicability of a widely-used basic task, the *n*-back task

(Kirchner, 1958), to a domain-specific task called the MRACT. This research extends an adaptive training system approach introduced in 2022 (Rebensky et al., 2022) and further delves into the connection between military and basic science tasks for evaluating competencies. Both tasks were implemented in a driving research adaptive testbed and were expected to similarly challenge working memory. The findings of this study reveal similarities between the two tasks but also points of substantial divergence. Physiological measures and environment measures showed value in extended from research tasks to more applied tasks. However, many variables were impacted by the tasks designs themselves. The findings, have implications for developing algorithms, adaptive training systems, and tools in controlled environments and their applicability to directly support the warfighter. This work contributes to discussions of developing adaptive algorithms and domain relevant tasks and has implications for practitioners looking to translate tasks to their operational setting. The discussion provides insights to inform the community on considerations and lessons learned to traverse the valley of death for technological solutions.

ACKNOWLEDGEMENTS

We would like to thank our engineers who created the DART testbed, Shawn Turk and Jonathan Reynolds. This work was supported by the Air Force Research Laboratory (AFRL) 711th Human Performance Wing (HPW/RHW) Gaming Research Integration for Learning Laboratory (GRILL) Contract Number: FA8650-21-C-6273. The views, opinions and/or findings are those of the authors and should not be construed as an official Department of the Air Force position, policy, or decision unless so designated by other documentation.

REFERENCES

Aleven, V., McLaughlin, E. A., Glenn, R. A., & Koedinger, K. R. (2016). Instruction based on adaptive learning technologies. *Handbook of research on learning and instruction*, *2*, 522-560.

Çankaya, S. (2019). Use of VR headsets in education: a systematic review study. Journal of Educational Technology & Online Learning, 2(1), 74-88.

Eddy, M. D., Hasselquist, L., Giles, G., Hayes, J. F., Howe, J., Rourke, J., Coyne, M., O'Donovan, M., Batty, J., Brunyé, T. T., & Mahoney, C. R. (2015). The Effects of Load Carriage and Physical Fatigue on Cognitive Performance. *PLOS ONE, 10*(7), e0130817. https://doi.org/10.1371/journal.pone.0130817

Giang, W., Santhakumaran, S., Masnavi, E., Glussich, D., Kline, J., Chui, F., Burns, C., Histon, J., & Zelek, J. (n.d.). Literature Review of Ecological Interface Design, Multimodal Perception and Attention, and Intelligent Adaptive Multimodal Interfaces. 269.

Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In *Advances in psychology* (Vol. 52, pp. 139-183). North-Holland.

Heff, C., Heger, D., Fortmann, O., Hennrich, J., Putze, F., & Schultz, T. (2014). Mental workload during n-back task—Quantified in the prefrontal cortex using fNIRSS. *Frontiers in Human Neuroscience, 7. https://doi.org/10.3389/fnhum.2013.00935*

Henrich, J., Heine, S. J., & Norenzayan, A. (2010). The weirdest people in the world?. *Behavioral and brain sciences*, *33*(2-3), 61-83.

Kelley, C. R. (1969). What is Adaptive Training? *Human Factors, 11*(6), 547–556. https://doi.org/10.1177/001872086901100602

Kirchner, W. K. (1958). Age differences in short-term retention of rapidly changing information. *Journal of experimental psychology*, *55*(4), 352.

Landsberg, C. R., Astwood Jr, R. S., Van Buskirk, W. L., Townsend, L. N., Steinhauser, N. B., & Mercado, A. D. (2012). Review of adaptive training system techniques. *Military Psychology, 24*(2), 96-113.

Lenné, M. G., Hoggan, B. L., Fidock, J., Stuart, G., & Aidman, E. (2014). The Impact of Auditory Task Complexity on Primary Task Performance in Military Land Vehicle Crew. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting, 58(1), 2185–2189. https://doi.org/10.1177/1541931214581459*

Lim, C., Barragan, J. A., Farrow, J. M., Wachs, J. P., Sundaram, C. P., & Yu, D. (2023). Physiological Metrics of Surgical Difficulty and Multi-Task Requirement during Robotic Surgery Skills. *Sensors*, *23*(9), 4354.

Matthews, G., Reinerman-Jones, L. E., Barber, D. J., & Abich IV, J. (2015). The psychometrics of mental workload: Multiple measures are sensitive but divergent. Human factors, 57(1), 125-143.

Mehler, B., Reimer, B., Coughlin, J. F., & Dusek, J. A. (2009). Impact of Incremental Increases in Cognitive Workload on Physiological Arousal and Performance in Young Adult Drivers. *Transportation Research Record: Journal of the Transportation Research Board, 2138(1), 6–12. https://doi.org/10.3141/2138-02*

Rebensky, S., Perry, S., & Bennett, W. (2022). How, when, and what to adapt: Effective adaptive training through game-based development technology. In *Interservice/Industry Training, Simulation, and Education Conference (I/ITSEC)*, Orlando, FL: I/ITSEC.

Schatz S, & Walcutt J. (2022). Modeling what matters: AI and the future of defense learning. *The Journal of Defense Modeling and Simulation*, *19(2),* 129-131.

Stalker, W., Rebensky. S., Knight, R., Perry, S., Bennett, W. (2024). The power of performance and physiological state: Approaches and considerations in adaptive game-based simulation. In *Adaptive Instructional Systems: 6th International Conference, AIS 2024, Held as Part of the 26th HCI International Conference, HCII 2024,* Washington, DC*,* HCII

Tole, J. R., Stephens, A. T., Harris Sr, R. L., & Ephrath, A. R. (1982). Visual scanning behavior and mental workload in aircraft pilots. *Aviation, Space, and Environmental Medicine*, *53*(1), 54-61.

Unni, A., Ihme, K., Surm, H., Weber, L., Ludtke, A., Nicklas, D., Jipp, M., & Rieger, J. W. (2015). Brain activity measured with fNIRSS for the prediction of cognitive workload. *2015 6th IEEE International Conference on Cognitive Infocommunications (CogInfoCom), 349–354. https://doi.org/10.1109/CogInfoCom.2015.7390617*

Vidulich, M. A., & Tsang, P. S. (2012). Mental workload and situation awareness. *Handbook of human factors and ergonomics*, 243-273.

Vine, C., Coakley, S., Myers, S.D., Blacker, S.D., & Runswick, O.R. (2020). The Reliability of a Military Specific Auditory n-Back Task and Shoot/Don't-Shoot Task. Psyarxiv.com/89vb5. 10.31234/osf.io/89vb5

Vine, C. A. J., Myers, S. D., Coakley, S. L., Blacker, S. D., & Runswick, O.R. (2021). Transferability of Military-Specific Cognitive Research to Military Training and *Operations. Front. Psychol., 12*(604803). 10.3389/fpsyg.2021.604803

Vygotsky, L. S. (1978). Mind in Society: The Development of Higher Psychological Processes. Cambridge, MA: Harvard University Press.

Zahabi, M., Park, J., Razak, A. M. A., & McDonald, A. D. (2020). Adaptive driving simulation-based training: framework, status, and needs. *Theoretical Issues in Ergonomics Science, 21*(5), 537–561. https://doi.org/10.1080/1463922X.2019.1698673