Enhancing a Piloting Task Simulator with Real-time Performance Feedback, Autopilot Disruption, Shock Punishment, and Adaptive Task Difficulty

Aaron Novstrup[1], Monica Tynan[2], Jonathan Lin[3], James Heaton[4]
Stottler Henke Associates[1], Wellman Center for Photomedicine[2], Massachusetts General Hospital Voice Center[34]
Seattle Washington[1], Boston Massachusetts[234]
anovstrup@stottlerhenke.com[1], mtynan2@partners.org[2], jonathanzlin@gmail.com[3],
James.Heaton@mgh.harvard.edu[4]

ABSTRACT

The computerized *multi-attribute task battery* (MATB) was introduced by NASA in the 1990s for cognitive workload research. This low-fidelity flight-simulation platform engages users in multiple tasks simultaneously, emulating representative piloting tasks, and is designed to engage multiple cognitive faculties at once (e.g., auditory/linguistic processing, visual processing, logical reasoning, visuo-motor coordination). The MATB has since been widely utilized in research investigating human performance, cognitive workload, trust in automation, strategic behavior, and related topics. Although it has been updated and re-implemented for a broader range of applications and greater researcher configurability by NASA, the U.S. Air Force, and others (e.g., MATB-II, AF-MATB, and OpenMATB, respectively), it still lacks key elements for maintaining participant engagement such as adequate real-time performance feedback and individualized adjustment of task difficulty.

This paper describes novel MATB enhancements designed for application in cognitive workload research and related domains. Specifically, the MATB was extended with real-time performance feedback, including auditory and haptic feedback and noxious electrical stimuli (i.e., shocks, individually calibrated to be unpleasant *but not painful*), designed to increase participant engagement, self-awareness, and motivation. Feedback of correct/incorrect responses and error punishment were expected to reduce the attenuation of physiological reactions to cognitive workload observed in simulated work environments relative to real ones. Additional MATB enhancements include periodic, instantaneous task-load self-assessment, an adaptive approach to manipulating task demands designed to improve experimental control for potentially confounding effects of learning and/or fatigue, and a novel task automation mode designed to bring “automation surprise” and/or “automation frustration” under experimental control. These modifications were tested in the context of a DARPA-supported human study (see Novstrup et al., 2023) investigating physiological responses to cognitive workload in human-automation teams (e.g., aircrew-automation teams). The paper concludes with observations and recommendations for researchers wishing to adopt these or similar mechanisms in their experimental designs.

ABOUT THE AUTHORS

Aaron Novstrup is an applied artificial intelligence researcher at Stottler Henke Associates, Inc., where he develops intelligent software systems, with a current focus on automated human state/performance assessment and monitoring.

Monica A. Tynan is a research technologist at the Wellman Center for Photomedicine at Massachusetts General Hospital (MGH) in Boston, MA. She received her B.S. in Biological Sciences from the University of Rhode Island in 2017 and her MSc in Applied Neuroscience from King's College London in 2022.

Jonathan Z. Lin completed his master's in Math and Computation at Harvard University before visiting as a research fellow at the MGH Voice Center. Jonathan was previously a systems analyst at MIT Lincoln Laboratory.

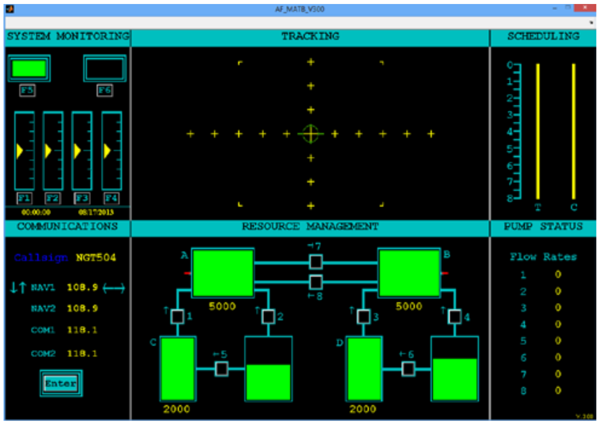
James T. Heaton, Ph.D. is Director of the Laryngeal Surgery Research Laboratory at MGH, Adjunct Professor at the MGH Institute of Health Professions, and Associate Professor of Surgery at Harvard Medical School. His research interests include voice and speech physiology, focusing on using electromyography for automatic speech recognition, cognitive workload assessment, and developing implantable stimulators to treat laryngeal paralysis.

Enhancing a Piloting Task Simulator with Real-time Performance Feedback, Autopilot Disruption, Shock Punishment, and Adaptive Task Difficulty

Aaron Novstrup[1], Monica Tynan[2], Jonathan Lin[3], James Heaton[4]
Stottler Henke Associates[1], Wellman Center for Photomedicine[2], Massachusetts General Hospital Voice Center[34]
Seattle Washington[1], Boston Massachusetts[234]
anovstrup@stottlerhenke.com[1], mtynan2@partners.org[2], jonathanzlin@gmail.com[3],
James.Heaton@mgh.harvard.edu[4]

BACKGROUND

NASA introduced the computerized multi-attribute task battery (MATB) for cognitive workload research in the 1990s (Comstock & Arnegard, 1992), and it has since been widely utilized in research investigating human performance, cognitive workload, trust in automation, strategic behavior, and related topics (e.g., Kong et al., 2022). The battery is implemented as a graphical software program and operated with a standard keyboard and joystick. It simultaneously engages multiple cognitive faculties (e.g., auditory/linguistic processing, visual processing, logical reasoning, visuo-motor coordination) with tasks designed to emulate representative piloting tasks. A screenshot of the MATB graphical user interface from the U.S. Air Force's AF-MATB implementation appears in Figure 1.



Callsign NGT504
NAV1 108.9
NAV2 108.9
COM1 118.1
COM2 118.1

Flow Rates

1	0
2	0
3	0
4	0
5	0
6	0
7	0
8	0

Figure 1. Screenshot of AF-MATB Graphical User Interface

Four cognitive tasks comprise the MATB: System Monitoring, Tracking, Communication, and Resource Management. The System Monitoring task (depicted in the top-left panel in Figure 1) involves monitoring a set of simulated lights and gauges and responding with an appropriate key or button press within a specified time limit if one of them deviates from the nominal state. The Tracking task (depicted in the top-center panel in Figure 1) involves using a joystick to steer a targeting reticle (cross-hairs) toward the center of the panel, in response to the reticle's scripted drift. The Communication task (depicted in the bottom-left panel in Figure 1) involves tuning one of four "radios" to a particular channel in response to a verbal prompt. The Resource Management, Pump Status, and Scheduling panels were not used in this study so that participants could achieve higher performance levels on a sub-set of MATB tasks during their single study visit.

The MATB has been updated and re-implemented for a broader range of applications and greater researcher configurability by NASA, the U.S. Air Force, and others (e.g., MATB-II, AF-MATB, and OpenMATB, respectively; (Cegarra et al., 2020; Comstock & Arnegard, 1992; Miller et al., 2014; Santiago-Espada et al., 2011). Existing implementations offer researchers substantial control over task parameters and the timing of simulation events, lending to the MATB's utility as an experimental instrument in several research areas. Despite their flexibility, critical limitations of the existing implementations reduce the instrument's applicability and, potentially, the generalizability of research results obtainable with the MATB. Within the context of DARPA-supported human studies investigating physiological responses to cognitive workload in human-automation teams (Piela, 2019; Urbano, 2021) the authors specifically encountered the following **limitations of the AF-MATB implementation** (v4.9):

- **A lack of performance feedback** left participants largely unaware of their performance and sometimes even their goals for each task despite verbal instruction and task practice/training. For example, gauges and lights could go into fault states and then return to normal after reaching a timeout without ever being noticed. Alternatively, participants could see the gauges and lights change from a fault state to a normal state and not know whether they timed out or were corrected by the simulated automation. The multiple task aspect of the MATB loses validity if participants are either unaware of task errors (due to lack of adequate feedback) or if they easily ignore tasks because their errors are not announced and there is no risk of consequence.
- **A lack of consequences for poor performance** may be a significant factor in the observed attenuation (see, for example, Angelborg-Thanderz, 1990) of human responses to similar task conditions in simulated environments relative to real ones. Attenuated responses are a particular concern in research seeking to relate cognitive workload or other internal states to observable physiological responses, where any attenuation increases the risk of Type II statistical errors (i.e., failure to detect an effect where one does, in fact, exist).
- **An inability to adjust task difficulty over time based on the skill or performance** of the participant may result in insufficient or inconsistent variation in the difficulty experienced. Variations in the relationship between task demands and experienced difficulty can occur over time due to learning, fatigue, boredom, etc. For example, the cognitive demands imposed by consistent task demands may diminish as a participant learns strategies and skills for performing the task. The scripting tools in existing MATB implementations provide means to specify a static set of predefined task difficulty levels, but not the means to adapt manipulations based on such dynamic effects.
- **Insufficient awareness of automation system behavior** prevents researchers from using the MATB to explore differences between human responses to automation failure and responses to more general task demands. Specifically, if participants cannot distinguish conditions imposed by simulated automation (and automation failures) from those that would exist in the absence of any task automation, eliciting a human response to automation failure is not feasible or meaningful.
- **A lack of rapid task-load self-assessment during MATB performance** misses an opportunity to document participants' subjective load during their performance. Current implementations of the MATB often assess task load using the National Aeronautics and Space Administration Task Load Index (NASA-TLX; Hart & Staveland, 1988) as a multi-dimensional scale collected after MATB trials (e.g., after task completion). However, instantaneous self-assessment of task difficulty obtained during MATB performance (or during short pauses in otherwise ongoing performance) could relate better to ongoing physiological signals and performance than more cumbersome self-assessment measures collected after the fact.
- **A lack of structured verbal responses** limits assessment of voice and speech production changes related to cognitive load. Several aspects of speech are known to change in the context of cognitive workload (Heaton et al., 2020; MacPherson, 2019; Quatieri et al., 2017) and sympathetic nervous system arousal (MacPherson, Abur, and Stepp, 2017; Dahl & Stepp, 2021). Current implementations of the MATB do not elicit verbal responses for acoustic assessment.

The following sections describe four specific enhancements developed and tested to address these limitations. First, the MATB was extended with real-time performance feedback, including auditory and haptic feedback and noxious electrical stimuli (i.e., shocks, individually calibrated to be unpleasant but not painful), designed to increase participant engagement, self-awareness, and motivation. Second, an adaptive approach to manipulating task demands was developed to improve experimental control for the potentially confounding effects of learning and/or fatigue. Third, a novel task automation mode was implemented to bring "automation surprise" or "automation frustration" under experimental control. Finally, a computerized implementation of the Instantaneous Self-Assessment of Workload (ISA) (Tattersall and Foord, 1996; Jordan and Brennen, 1992) was incorporated into the MATB whereby participants verbally reported task difficulty at the end of a structured phrase. All four enhancements were developed as modifications to the Air Force MATB software implementation (AF-MATB v4.9) and tested within the context of a human study of physiological responses to cognitive workload in human-automation teams (e.g., aircrew-automation teams). Pertinent details of the companion human study are introduced in this paper only as necessary, while more complete information can be found in Novstrup et al. (2023), which reports on the primary results of the authors' most recent human study. The present paper focuses only on the MATB enhancements developed to support that study.

REAL-TIME PERFORMANCE FEEDBACK

The AF-MATB was extended to provide participants with real-time feedback for their task performance. Participants were provided with positive feedback for correct/adequate task performance (i.e., responding with the correct actions within established time limits) and negative feedback for incorrect/inadequate task performance (i.e., failing to respond within specified time limits or responding with incorrect actions). Two primary stimulus modalities were used to provide performance feedback: acoustic and electrical. A form of mildly assistive haptic performance feedback was also provided by a new automation mode implemented with a force feedback joystick—automatically pushing the joystick to steer toward the center of the targeting pane when the reticle drifted beyond a pre-determined threshold. The Automation Surprise Manipulation section below describes this automation mode in more detail. Feedback of correct/incorrect responses and error punishment were expected to reduce the attenuation of physiological reactions to cognitive workload observed in simulated work environments (e.g., Angelborg-Thanderz, 1990).

Acoustic feedback consisted of various sound effects and verbal cues. Correct responses in the System Monitoring and Communication tasks were "rewarded" with a pleasant chime sound (a 'ding' commonly associated with task completion on personal computers). Incorrect keystrokes for System Monitoring were "punished" with a telephone-off-the-hook sound, while System Monitoring timeouts were punished with a buzzer sound. Tracking out-of-range errors, corresponding to drift of the Tracking reticle beyond a predetermined limit, were also punished with a buzzer sound distinct from other sounds and repeatedly played while the reticle remained out of range. Incorrect responses in the Communication task were punished with a different sound followed by a verbal "radio error" announcement. Finally, timeouts in both the System Monitoring and Communication tasks were indicated by an unpleasant buzzer sound followed by a verbal announcement of "lights", "gauges", or "communication" as appropriate.

In addition to acoustic feedback, noxious electrical stimuli (i.e., shocks) were used as error consequences with methods similar to those detailed in Lindström et al., (2013). Brief (100 ms), unpleasant electrical shocks were administered after task errors, 200 ms after acoustic error feedback, using the following methods.

- Monopolar electrical pulses were administered to the forearm skin at voltages ranging from approximately 30–70V (see Lindström et al., 2013). The voltage was individually calibrated to one that participants reported as *unpleasant but not painful*, using an adaptive staircase procedure (Treutwein, 1995) in 1–5V steps. Starting from 25V, an experimenter iteratively administered a 100 ms pulse and asked the subject to rate the sensation using an 11-point pictographic faces pain scale (Figure 2 ; Hicks et al., 2001). The calibration process terminated at an estimate of the highest voltage the subject experienced as unpleasant but not painful.

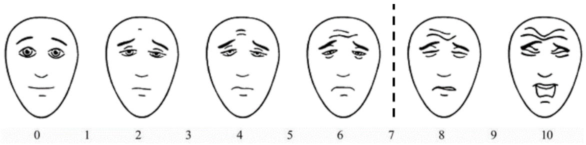


Figure 2. Faces Pain Scale (Hicks et al., 2001)

- Shocks were delivered using two Ag/AgCl disposable 20 mm diameter circular electrodes (Natus Medical) placed on the left dorsolateral forearm skin approximately 5 cm apart. Shock voltage was controlled by a constant-voltage stimulation module (BIOPAC STM200; see Figure 3) and delivered current was monitored by a BIOPAC MP160 data acquisition system.
- To limit the number of shocks delivered, the periods of task engagement were sub-divided in advance into 20-second intervals in which, at most, one shock could occur regardless of the number of errors made. This scheme allowed two shocks to occur nearly back-to-back if they happened to fall at the end of one such interval and at the beginning of the next, but prevented three or more shocks from occurring in rapid succession and strictly limited the maximum number of shocks administered in each period of task engagement to the number of non-overlapping 20-second intervals in each such period.

Although AF-MATB features verbal prompts for the Communication task, its acoustic playback system is not designed for concurrent playback of multiple sounds, which can cause software instability and early termination (i.e., crashing). Therefore, both acoustic and electrical feedback were implemented with AF-MATB's digital-port-triggering feature, a customized software-to-hardware bridge, and an external soundboard (WAV Trigger by Robersonic.com; Figure 3). The digital-port-triggering feature is typically used to synchronize scripted MATB events with external signals (e.g., for data acquisition). A corresponding event code is transmitted on a digital communication port when a scripted event occurs. Here, the AF-MATB component responsible for monitoring participant performance was modified to emit event codes associated with key performance-related events (e.g., task errors). The open-source com0com software null-modem emulator was used to deliver the event codes to a separate software process (implemented in Python) that translated the event codes into digital signals and transmitted them to the soundboard via a digital I/O board (National Instruments USB6501; Figure 3). The Python software was also responsible for enforcing the limit on the number of electrical stimuli (i.e., ≤1 per 20s increment of MATB testing). Stimulation voltage and pulse width were controlled and limited (respectively) by the stand-alone stimulator (STM200), ensuring that possible software errors could not result in harmful shocks.

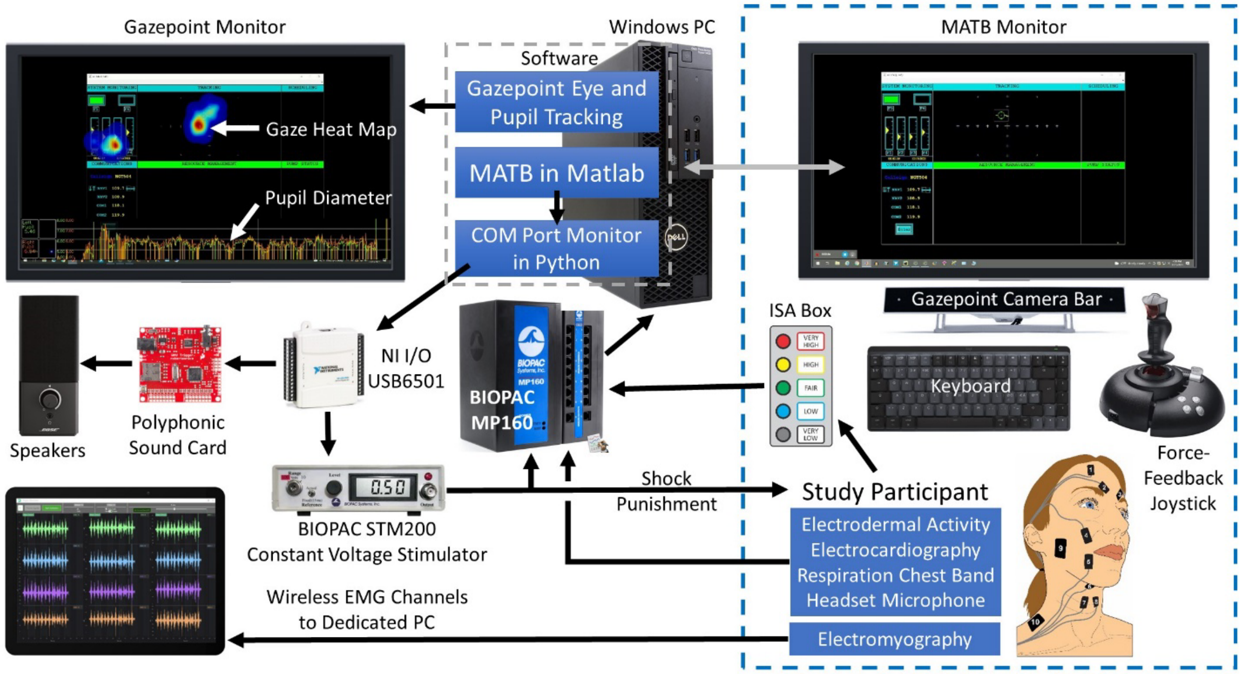


Figure 3. Enhanced MATB System Diagram with Performance Feedback (via Sounds and Shocks), Instantaneous Self-Assessment (ISA) Reporting, and Physiological Signal Recording. A blue dotted box highlights the MATB-human interface.

The polyphonic soundboard was preloaded with a set of sound files (16-bit, 1411kbps) for each as acoustic stimuli and a file containing a 100ms square waveform for triggering shock punishment. It mapped each digital signal emitted by the digital I/O board to the appropriate sound file and analog output channel with a short 7ms trigger-to-sound delay and allowed simultaneous playback of audio files. One output channel was wired to speakers for acoustic output, and one was wired to the STM200 for triggering electrical stimuli. The constant-voltage output from the STM200 passed through a current feedback monitor cable (BIOPAC CBLCFMA) attached to the MP160 through an electrical isolation module to monitor and record the stimulation current.

We previously introduced the use of noxious shock punishment in relation to MATB performance to increase cognitive load by heightening concern over task performance (Piela, 2019; Urbano, 2021). In those works, testing trials were divided equally into "Low Consequence" and "High Consequence" conditions, similar to methods employed by Lindström et al., 2013. Participants did not receive shock punishment during the Low Consequence trials relating to performance errors. For High Consequence trials, participants anticipated receiving between 1-5 shocks immediately after the trial completion in proportion to task errors. A seventh subscale of "concern" regarding MATB performance was added to the NASA-TLX that was collected between trials before shock delivery to measure the psychological impact of shock punishment threat. As anticipated, the threat of shocks significantly increased study participants' reported concern over trial performance (Piela, 2019; Urbano, 2021). In addition, mean electrodermal activity was higher during High Consequence trials than Low Consequence trials, consistent with heightened sympathetic nervous system output when threatened with shock punishment (Urbano, 2021). We therefore concluded that the High

Consequence condition provides the best context for studying cognitive workload, motivating us to incorporate that form of punishment/feedback in all trials for recent experiments (Novstrup et al., 2023). Moreover, behavioral research has historically shown that such feedback is most effective when accurate and paired closely in time (Brand et al., 2020; Garris et al., 2002), so shock delivery was changed from after trial completion to immediately upon error commission (with restrictions – see previous description) in our most recent set of MATB modifications.

ADAPTIVE TASK DEMANDS

Existing MATB implementations provide the means for experimenters to control MATB task demands by scripting the occurrence of discrete events and/or specifying simulation control parameters. For example, in the AF-MATB implementation, the demands of the Tracking task are controlled through the specification of parameters that determine the speed of the targeting reticle's drift and the frequency with which the drift direction changes. In an AF-MATB script, the experimenter specifies these Tracking difficulty parameters for three distinct Tracking difficulty levels (i.e., "low", "medium", and "high") and must associate every time period in the script with one of these levels. Thus, the task conditions are determined in advance and cannot be adjusted dynamically based on the observed performance of the participant.

The authors' companion study (Novstrup et al., 2023) identified contrasts between physiological responses to challenging yet manageable cognitive demands versus cognitively overloading task demands. Probing the boundary between such conditions efficiently, even in the presence of dynamic effects such as learning and fatigue, required a means to tailor task demands to individual performance. Several modifications were made to the AF-MATB implementation to make this possible. First, a holistic (i.e., multi-task) performance scoring mechanism was introduced to capture the performance criteria by which task demands could be dynamically adjusted. Second, the means were introduced to override specific AF-MATB script parameters upon script execution. Finally, rules were implemented for dynamically setting task difficulty parameters based on a combination of scripted parameters, parameter value overrides specified upon script execution, and performance scores observed during the execution of a script. These modifications are described in further detail below.

A performance score ranging from 0 to 100 was introduced to capture a holistic measure of task performance. Since the Tracking task was taken to be the "primary" task in the study (although participants were instructed to perform all three enabled tasks to the best of their abilities) and all task demand manipulations were carried out through Tracking task conditions, the performance score was based on Tracking task performance and then discounted for errors on the other tasks. More specifically, a participant's performance score over a given period was simply the percentage of time during which the Tracking reticle was within a predetermined distance from the center of the Tracking panel, with fixed amounts deducted for each incorrect response or timeout on the System Monitoring or Communication task and then any negative result replaced with a score of zero. The AF-MATB software component responsible for monitoring performance-related events was modified to maintain the information necessary to compute the performance score. The software was also modified to compute the score and report this information after each period of task engagement.

To facilitate the dynamic adjustment of task difficulty across multiple AF-MATB executions, the software was modified to report, upon completing script execution, the (dynamically determined) Tracking task difficulty parameters for the most difficult period of task engagement in which the participant scored above a configured threshold—96 points in the authors' companion study (2023). This represented a good but non-perfect performance that could serve as the starting point for the next script (e.g., the next MATB testing trial). The software was also modified to prompt the experimenter for an optional set of difficulty parameters immediately before script execution. This set of parameters overrode the scripted "low" difficulty parameters for the Tracking task based on performance of the prior script. Thus, this mechanism enabled task performance in previous AF-MATB executions to influence task difficulty in later executions (see Figure 4).

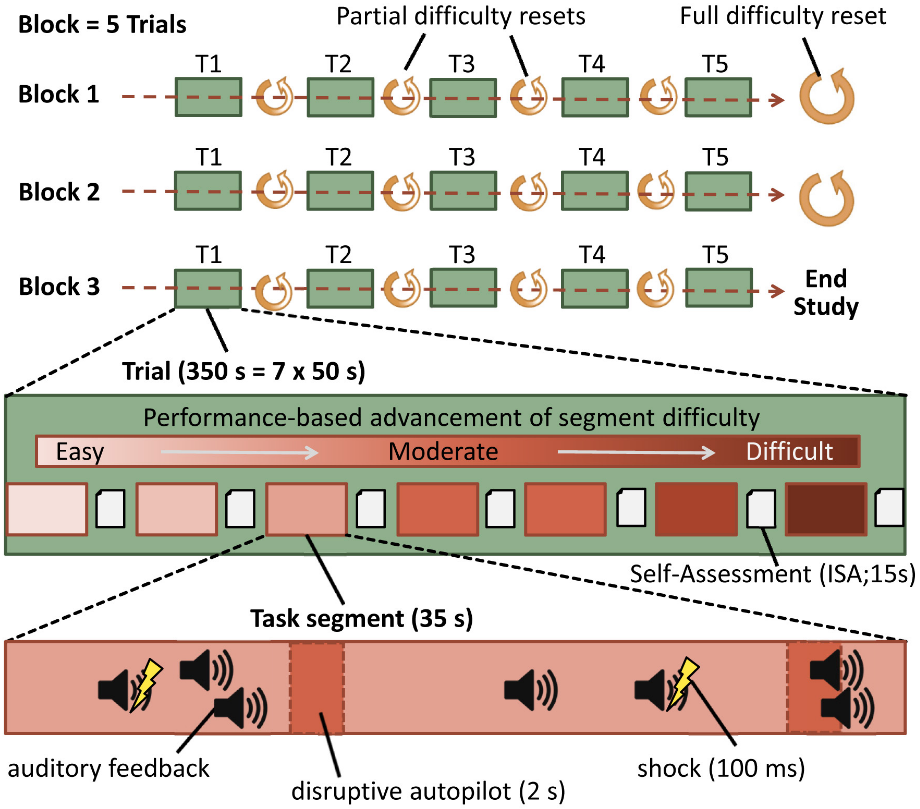


Figure 4. Recording Session Design (Novstrup et al., 2023)

The AF-MATB script interpreter was altered to determine Tracking task difficulty parameters as a function of their scripted values, optional override values specified upon script execution, and performance scores observed during execution of the script. This function assigned the values by linearly interpolating between those for the "low" and "high" difficulty levels, with an interpolation parameter that was a function of performance scores for each of the previous periods of task engagement in the script. In the authors' study specifically, each script defined seven, 50s periods or segments of task engagement, starting at the "low" difficulty level (as defined by either the overrides or by the scripted values if no overrides were specified), ratcheting up after each segment in which the participant achieved a performance score of at least 92 points, and reaching the "high" difficulty level in the seventh segment if and only if the difficulty increased in every segment. Therefore, Tracking task difficulty increased from one segment to the next when performance was at least 92 points in the previous segment, and the starting difficulty of the next trial was set at the highest level where at least 96 points were achieved in a segment within the previous trial (Figure 4).

With this approach, Tracking difficulty does not get easier across segments (within trials) or from trial to trial within a block of trials, regardless of performance. In turn, performance can be expected to decrease across trials within each block and across segments within each trial due to performance-based Tracking difficulty escalation as described above. This pattern was indeed observed in the authors' study, with performance dropping across both trials and segments (Figure 5). However, in the authors' study, Tracking difficulty was reset to the easiest level after each block of 5 trials to restore cognitive load toward pre-escalation baseline. Consequently, the first trial of each block (trials 1, 5, 10) after each difficulty reset yielded the best performance, the second trial in each block (trials 2, 6, 11) yielded slightly lower performance, and this pattern continued. After the initial 3-4 segments of each trial, performance became more similar across trial averages. Tracking task difficulty had already escalated for each participant based on their prior performance, thereby challenging everyone towards their performance plateau by later segments (see Figure 5). Learning across blocks of trials was evident in the authors' study, with average performance being better in blocks 2 and 3 than block 1 (see Figure 5, right plot). As desired, however, the impact of learning was small relative to the influence of manipulated task difficulty across segments.

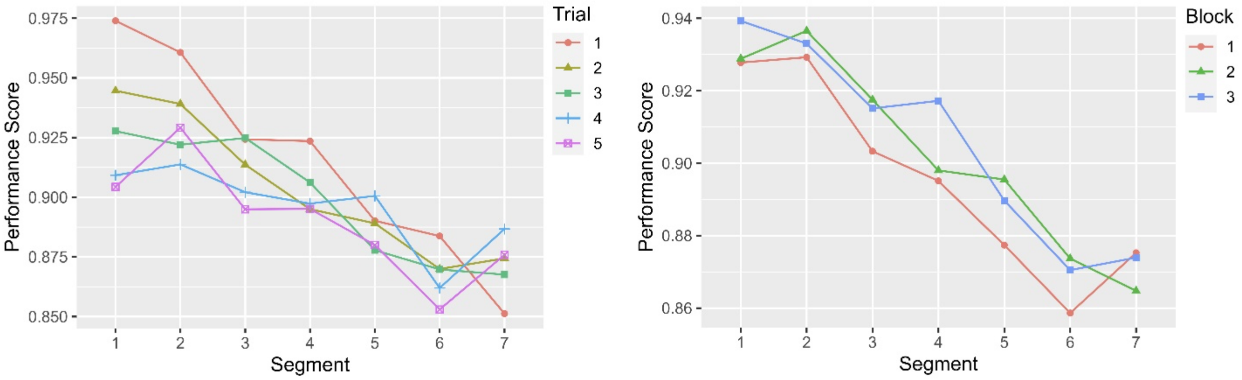


Figure 5. Performance on MATB Across Trial Segments by Trials (Left Plot) and by Blocks (Right Plot)

AUTOMATION SURPRISE MANIPULATION

A key motivation of the authors' study was to identify physiological responses to cognitive states of "automation surprise" or "automation frustration" instead of more generalized cognitive states associated with other sources of variation in cognitive task demands. We define these hypothetical cognitive states as ones that may arise when an automation system behaves correctly yet in a way that a human team member does not expect (automation surprise) or increases task demands for a human team member by behaving incorrectly (automation frustration). Both states require human awareness of the automation system's incorrect or unexpected behavior, and therefore assume an ability of a participant to distinguish conditions imposed by the behavior (or failure) of the automation system from those that result from other factors in the task environment.

Although existing implementations of MATB provide simulated automation modes (including failure modes), it is frequently difficult for participants to distinguish automation failure modes from ordinary task conditions. For example, the AF-MATB provides an automation mode for the System Monitoring task in which the simulated automation detects a light or gauge fault and corrects it after a brief period or, in its failure mode, fails to detect the fault state. Since light/gauge fault states are always eventually corrected when a timeout occurs, a participant may not even notice a failure of this automation mode. Some implementations of the Tracking task provide a simulated automation mode in which the targeting reticle drifts within a small radius of the central target, with a simulated failure mode identical to the conditions in the ordinary Tracking task. The difficulty of identifying such simulated automation failures as currently offered makes them unreliable sources of stimuli for inducing states of automation surprise/frustration. Therefore, the authors developed a new automation mode for the Tracking task to (it was hoped) more reliably induce such states.

The new automation mode was implemented as a simulated autopilot for the Tracking task (Figure 6). The autopilot was operationalized as a software controller for a force feedback joystick (the Microsoft Sidewinder Force Feedback 2), implemented within the AF-MATB source code. Unlike a standard joystick, a force feedback joystick is both an input and output device, capable of actively applying forces to the control stick (typically with two internal servomotors) under the direction of software. The Microsoft Sidewinder Force Feedback 2 joystick is a commercial product designed primarily for application in video gaming, where it can provide various forms of haptic feedback to enhance the realism and immersiveness of the gaming experience. Its adherence to open interface standards for gaming controllers and compatible force feedback support in MATLAB (the language and software runtime environment in which AF-MATB is implemented) made it relatively straightforward to integrate with AF-MATB.

The simulated autopilot was developed with two modes under experimental control: an assistive mode and a disruptive mode. The assistive mode engaged whenever the Tracking reticle strayed beyond a configured range and disengaged again after the reticle had remained in range for a configured duration. When the assistive autopilot mode was engaged, the software controller induced forces on the joystick, tending to drive it from its position toward a position that would steer the targeting reticle toward the central target. Thus, this mode provided a mildly assistive form of haptic feedback for performance in the Tracking task. The disruptive autopilot mode engaged in scripted intervals and automatically disengaged the assistive mode when applicable. This mode acted as a simulated automation failure mode by inducing forces on the joystick that tended to align the joystick position with the reticle drift vector (i.e., amplifying the scripted drift) while also causing the joystick to shake violently.

Both autopilot modes were implemented as a controller that updated the two force parameters (one for each of the joystick's two axes of motion) assigned to the force feedback joystick within the AF-MATB's 10 Hz event loop. In each cycle, the controller first computed a target position for the joystick based on the autopilot mode, the position of the targeting reticle, and the reticle drift vector. In the assistive mode, the target joystick position would steer toward the central target after compensating for the magnitude and direction of reticle drift. In the disruptive mode, the target joystick position would steer the joystick in the same direction as the reticle drift, thereby amplifying the drift if not physically counteracted by the operator. Next, the controller computed the difference between the joystick and target joystick positions. This vector determined the primary component of the force vector assigned as the joystick's force parameters. In the disruptive mode, this primary vector component was supplemented by a perpendicular component with an oscillating magnitude that resulted in violent shaking (while still steering, on average, in the direction of the reticle drift).

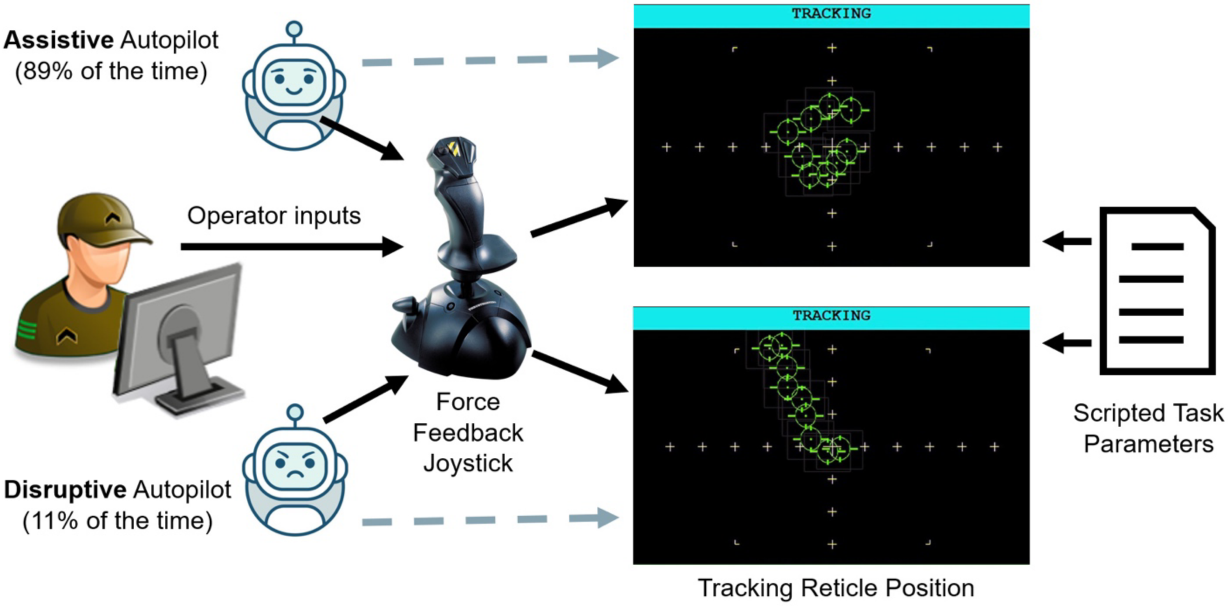


Figure 6. Operationalization of the Simulated Autopilot with a Disruptive Failure Mode. The operator and simulated autopilot co-determine joystick position. Scripted task parameters and joystick position in turn co-determine the movement of the targeting reticle in the Tracking task.

When simulated autopilot switched from assistive mode to disruptive mode in the authors' study (Novstrup et al., 2023), participants showed a conspicuous increase in ipsilateral skin surface electromyography (sEMG) magnitude of 5 microvolts (on average) from an arm muscle contributing to wrist extension (extensor carpi radialis longus) and, for about half of participants, a smaller (yet still significant) average increase in EMG magnitude from a shoulder muscle (the trapezius), as would be expected when the tracking task becomes physically more demanding. In addition, other face- and neck-surface sEMG recording locations showed statistically significant (α=0.05) increases in signal strength for at least half of the participants, including locations associated with positioning the eyebrows (corrugator supercilii), mouth/lips (e.g., zygomaticus major, orbicularis oris inferior), tongue (anterior digastric, mylohyoid), and extrinsic larynx (e.g., sternohyoid). This indicates that disruptive joystick movements were not only being actively counteracted by participants using arm and shoulder muscles but that users also reacted more broadly with stronger contractions from emotionally expressive muscles, which are theoretically independent of joystick control. Increased sEMG magnitude beyond what is needed to perform the Tracking task does not unequivocally identify changes in cognitive or emotional states, but increased contraction of the corrugator supercilii (Brown, 2019; Elkins-Brown et al., 2016; Larsen et al., 2003; Lindström et al., 2013) is often associated with stress and negative emotional valence (respectively), consistent with the disruptive autopilot working as intended.

INSTANTANEOUS SELF-ASSESSMENT OF WORKLOAD

Subjective workload is often assessed for the MATB using the NASA-TLX (for the NASA-MATB and AF-MATB) and other arbitrary rating scales similar to the TLX instrument (for OpenMATB). The TLX characterizes workload on six sub-scales, including mental, physical, and temporal demands, performance, frustration, and effort (Hart, 2006; Hart & Staveland, 1988). The purported advantage of capturing these sub-scales is that it increases the likelihood of representing at least one aspect of how each person conceptualizes workload (Casner & Gore, 2010). In addition to evaluating the workload sub-scales individually, the sub-scales can be rank-ordered based on the subjective importance of each element (for each participant) to generate a weighted average for a personalized, overall workload measure. Although the TLX can be obtained immediately after MATB trials while the experience is still somewhat fresh and reportable, it is relatively more time-consuming than other workload assessment techniques and is biased towards reporting what the user experienced immediately before reporting.

An alternative assessment technique that can be employed with the MATB is Instantaneous Self-Assessment (ISA) (Jordan & Brennen, 1992). ISA can take many forms, but the general notion is to probe the participant during, or during a brief pause in, an ongoing task. This is often done with a simple scale from 1 (low) to 5 (high), using descriptions of task difficulty appropriate for the task(s) at hand. In the authors' study using the modified AF-MATB platform, ISA was obtained during 15s pauses occurring at the end of every 50s trial segment, during which participants reported task difficulty as Very Low, Low, Fair, High, or Very High. Participants were prompted at each pause by a recording of "Please report task difficulty". They would respond by saying, "This is NGT504, task difficulty is _______", choosing among the five discrete difficulty levels. They also simultaneously confirmed task difficulty by pressing one of five buttons on a custom ISA Unit labeled with the five discrete difficulty levels (following the methods of Jordan & Brennen, 1992), producing an identification signal recorded along with the complement of AF-MATB and physiological data (see Figure 3).

An advantage of periodic ISA during MATB trials is the ability to obtain subjective task difficulty with minimal intrusion on MATB tasks, capturing changes in perceived difficulty within trials as Tracking difficulty changed in a performance-related manner. In the authors' study, self-reported task difficulty was positively correlated with Tracking task difficulty (Pearson's $r \geq 0.333$, $p < 0.0006$ for all participants), showing a consistent and intuitive relationship between scripted task difficulty and perceived task difficulty. Prior work has demonstrated that ISA can be disruptive to ongoing task performance (Tattersall & Foord, 1996), motivating our 15s pause in MATB tasks during ISA reporting. An alternative approach could have been to use only verbal or push-button responses rather than both. However, another motivation for using verbal responses was to obtain speech for assessing acoustic correlates of workload (Dahl & Stepp, 2021; Heaton et al., 2020; MacPherson, 2019; MacPherson et al., 2017; Quatieri et al., 2017), which was accomplished during the ISA pauses without competing joystick noise, sound file playback, etc.

DISCUSSION

Performance feedback is critical to skill acquisition — particularly with novel tasks where errors may go unnoticed and/or when task objectives may be poorly understood. Our initial experience with using the AF-MATB to study frustration (Piela, 2019) and cognitive load (Urbano, 2021) drew into question whether novice or inexperienced study participants recognized when they made errors when performing System Monitoring, Communications, and Tracking tasks, motivating our addition of performance feedback to the AF-MATB. Additional studies are needed to compare MATB performance with and without these simulation enhancements. Still, our initial experience with the modified platform (Novstrup et al., 2023) indicates that these enhancements increase participant engagement and understanding of task demands. Participants could associate the sounds for both correct and incorrect responses for the System Monitoring and Communication tasks, or reticle position for the Tracking task, rapidly and intuitively. This allowed them to perform relatively well in early trials and blocks (Figure 5), thus avoiding the need to drop initial trials due to performance variability, as was done in our prior MATB testing (Piela, 2019). In addition, although participants reported greater concern over their performance on trials with shock punishment for errors (Piela, 2019; Urbano, 2021), they seemed to embrace the added challenge of shock consequence for errors in our recent study (Novstrup et al., 2023). Not only did everyone complete the testing with real-time shock punishment (N=11), but forty-five percent of participants also requested one or more *increases* in shock voltage over their testing trials. In contrast, only eighteen percent of participants requested a voltage decrease after the initial shock calibration. These forms of continuous (auditory) or intermittent (shock) feedback potentially address some concerns over a lack of error consequence and system understanding for the MATB platform, as has been expressed for simulation testing environments lacking adequate feedback and situational awareness (Gouraud et al., 2017).

In addition to the added performance feedback and error consequence, participants were prompted to provide ISA of task difficulty every minute during MATB trials. ISA values correlated with the Tracking task difficulty in the authors' study (Novstrup et al., 2023), indicating that this was a valuable metric for assessing task difficulty perception on a minute-by-minute basis as Tracking difficulty levels changed. It is impractical to administer the NASA-TLX instrument (integrated into all existing MATB implementations) this frequently due to its relative complexity. To obtain the most information with the least amount of simulation intrusion, future work should compare the verbal and push-button response mechanisms to other means of administering ISA, and the frequency of ISA administration should also be examined (Jordan & Brennen, 1992; Tattersall & Foord, 1996).

Although they are correlated, the self-reported difficulty of a given task is subject to variation based on multiple factors other than just the task conditions themselves, including participant skill. Therefore, experimenters who aim to manipulate the subjective experience of difficulty may consider adapting task conditions based on participants' observed performance. The adaptive mechanism described here enabled the authors to present participants with a reasonably consistent range of challenges throughout task engagement, even as participants acquired skill in the tasks. This made it possible to efficiently probe for contrasts between responses to challenging yet manageable task loads and responses to cognitive overload. By better matching challenges to participants' skill, this (or a similar) mechanism may also promote a sense of "flow" (Jin, 2011), leading to greater participant engagement, motivation, and performance. Interpolating between a minimum difficulty level determined by observed past performance and a predetermined maximum difficulty level made the mechanism simple to implement and integrate with AF-MATB (i.e., by applying the interpolation in AF-MATB's difficulty parameter space). However, our implementation of adaptive task demands to date has only manipulated the Tracking task difficult, and with only a single set of performance values for increasing task difficulty across trial segments and across trials within testing blocks. Manipulating the difficulty of multiple tasks in relation to systematic manipulation of performance cut-off scores is needed in future experiments to explore the full potential usefulness of this MATB modification.

Finally, the new Tracking task automation described here provides a means to study psychophysiological responses to disruptive automation behavior. Unlike simulated automation modes available in existing MATB implementations, the behavior of the automation was clearly discernible from other aspects of the task environment. Data collected by Novstrup et al. (2023) demonstrated a significant difference in motor activity between the periods immediately before and after the onset of the disruptive autopilot mode. These changes in sEMG during disruptive autopilot periods were not only found in arm muscles but also in muscles of facial expression, indicating not only increased physical exertion associated with combating the disruptive effects of the automation but also increased cognitive stress and/or negative emotional responses consistent with automation surprise/frustration. Future studies should implement the disruptive autopilot across a range of force levels, disruption duration, and disruption frequency to identify the best way to elicit desired psychophysiological responses to automation failure.

ACKNOWLEDGEMENTS

The authors wish to thank the Air Force Research Laboratory (AFRL) and, in particular, Dr. Justin Estepp for their role in developing the AF-MATB software implementation and for providing its source code to DARPA for use in this research. This material is based upon work supported by the Defense Advanced Research Projects Agency (DARPA) under Contract No. 140D6318C0039[1]. The views, opinions, and/or findings expressed are those of the authors and should not be interpreted as representing the official views or policies of the Department of Defense or the U.S. Government.

[1] Approved for Public Release, Distribution Unlimited

REFERENCES

Angelborg-Thanderz, M. (1990). Military flight training at a reasonable price and risk. *Economics Research Institute, Stockholm School of Economics and FOA Report C*, 50083-5.1.

Brand, D., Novak, M. D., DiGennaro Reed, F. D., & Tortolero, S. A. (2020). Examining the Effects of Feedback Accuracy and Timing on Skill Acquisition. *Journal of Organizational Behavior Management*, *40*(1–2), 3–18. https://doi.org/10.1080/01608061.2020.1715319

Brown, N. (2019). *The Properties and Functional Significance of Error-Related Electromyographic Activity over the Corrugator Supercilii for the Study of Error Monitoring, Error Awareness, and Race* [Ph.D., University of Toronto (Canada)]. https://www.proquest.com/docview/2316839417/abstract/B9C570D525074DBCPQ/1

Casner, S. M., & Gore, B. F. (2010). Measuring and evaluating workload: A primer. *NASA Technical Memorandum*, *216395*, 2010.

Cegarra, J., Valéry, B., Avril, E., Calmettes, C., & Navarro, J. (2020). OpenMATB: A Multi-Attribute Task Battery promoting task customization, software extensibility and experiment replicability. *Behavior Research Methods*, *52*(5), 1980–1990. https://doi.org/10.3758/s13428-020-01364-w

Comstock, J. R., & Arnegard, R. J. (1992). *The multi-attribute task battery for human operator workload and strategic behavior research* (NAS 1.15:104174). https://ntrs.nasa.gov/citations/19920007912

Dahl, K. L., & Stepp, C. E. (2021). Changes in Relative Fundamental Frequency Under Increased Cognitive Load in Individuals With Healthy Voices. *Journal of Speech, Language, and Hearing Research*, *64*(4), 1189–1196. https://doi.org/10.1044/2021_JSLHR-20-00134

Elkins-Brown, N., Saunders, B., & Inzlicht, M. (2016). Error-related electromyographic activity over the corrugator supercilii is associated with neural performance monitoring. *Psychophysiology*, *53*(2), 159–170. https://doi.org/10.1111/psyp.12556

Garris, R., Ahlers, R., & Driskell, J. E. (2002). Games, Motivation, and Learning: A Research and Practice Model. *Simulation & Gaming*, *33*(4), 441–467. https://doi.org/10.1177/1046878102238607

Gouraud, J., Delorme, A., & Berberian, B. (2017). Autopilot, Mind Wandering, and the Out of the Loop Performance Problem. *Frontiers in Neuroscience*. https://doi.org/10.3389/fnins.2017.00541

Hart, S. G. (2006). Nasa-Task Load Index (Nasa-TLX); 20 Years Later. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, *50*(9), 904–908. https://doi.org/10.1177/154193120605000909

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). Elsevier.

Heaton, K. J., Williamson, J. R., Lammert, A. C., Finkelstein, K. R., Haven, C. C., Sturim, D., Smalt, C. J., & Quatieri, T. F. (2020). Predicting changes in performance due to cognitive fatigue: A multimodal approach based on speech motor coordination and electrodermal activity. *Clinical Neuropsychologist*, *34*(6), 1190–1214. https://doi.org/10.1080/13854046.2020.1787522

Hicks, C. L., von Baeyer, C. L., Spafford, P. A., van Korlaar, I., & Goodenough, B. (2001). The Faces Pain Scale – Revised: Toward a common metric in pediatric pain measurement. *Pain*, *93*(2), 173–183. https://doi.org/10.1016/S0304-3959(01)00314-1

Jordan, C. S., & Brennen, S. D. (1992). Instantaneous self-assessment of workload technique (ISA). *Defence Research Agency, Portsmouth*.

Kong, Y., Posada-Quintero, H. F., Gever, D., Bonacci, L., Chon, K. H., & Bolkhovsky, J. (2022). Multi-Attribute Task Battery configuration to effectively assess pilot performance deterioration during prolonged wakefulness. *Informatics in Medicine Unlocked*, *28*, 100822. https://doi.org/10.1016/j.imu.2021.100822

Larsen, J. T., Norris, C. J., & Cacioppo, J. T. (2003). Effects of positive and negative affect on electromyographic activity over zygomaticus major and corrugator supercilii. *Psychophysiology*, *40*(5), 776–785. https://doi.org/10.1111/1469-8986.00078

Lindström, B. R., Mattsson-Mårn, I. B., Golkar, A., & Olsson, A. (2013). In your face: Risk of punishment enhances cognitive control and error-related activity in the corrugator supercilii muscle. *PLOS One*, *8*(6), e65692.

MacPherson, M. K. (2019). Cognitive Load Affects Speech Motor Performance Differently in Older and Younger Adults. *Journal of Speech, Language, and Hearing Research*, *62*(5), 1258–1277. https://doi.org/10.1044/2018_JSLHR-S-17-0222

MacPherson, M. K., Abur, D., & Stepp, C. E. (2017). Acoustic Measures of Voice and Physiologic Measures of Autonomic Arousal during Speech as a Function of Cognitive Load. *Journal of Voice*, *31*(4), 504.e1-504.e9. https://doi.org/10.1016/j.jvoice.2016.10.021

Miller, J., Schmidt, K. D., Estepp, J. R., Bowers, M., & Davis, I. (2014). *An Updated Version of the U.S. Air Force Multi-Attribute Task Battery (AF-MATB)*.

Novstrup, A., Tynan, M. A., Kline, J., Deluca, G., & Heaton, J. T. (2023, November 27). *Towards robust estimation of cognitive workload from wearable physiological sensors*. I/ITSEC, Orlando Florida.

Piela, M. (2019). *Speech acoustic and other physiological correlates of frustration during human-computer interaction* [Masters of Science in Speech Language Pathology]. MGH Institute of Health Professions.

Quatieri, T. F., Williamson, J. R., Smalt, C. J., Perricone, J., Patel, T., Brattain, L., Helfer, B., Mehta, D., Palmer, J., & Heaton, K. (2017). Multimodal biomarkers to discriminate cognitive state. *The Role of Technology in Clinical Neuropsychology*, *409*.

Santiago-Espada, Y., Myer, R. R., Latorella, K. A., & Comstock, J. R. (2011). *The Multi-Attribute Task Battery II (MATB-II) Software for Human Performance and Workload Research: A User's Guide.* https://ntrs.nasa.gov/citations/20110014456

Tattersall, A. J., & Foord, P. S. (1996). An experimental evaluation of instantaneous self-assessment as a measure of workload. *Ergonomics*, *39*(5), 740–748. https://doi.org/10.1080/00140139608964495

Treutwein, B. (1995). Adaptive psychophysical procedures. *Vision Research*, *35*(17), 2503–2522. https://doi.org/10.1016/0042-6989(95)00016-X

Urbano, M. (2021). *Speech acoustic and other physiological correlates of cognitive load during human-computer interaction* [Masters of Science in Speech Language Pathology]. MGH Institute of Health Professions.