Making MATB-II medical: Pilot testing results to determine a novel lab-based, stress-inducing task

Lauren Kennedy
Virginia Polytechnic Institute and State University
Department of Translational Biology, Medicine, and Health
Blacksburg, VA, USA

Sarah Henrickson Parker, PhD
Virginia Tech Carilion School of Medicine
Department of Biomedical Sciences
Roanoke, VA, USA

The purpose of this project was to adapt an existing computer-based task called Multi-Attribute Task Battery (MATB-II), developed by NASA and frequently used to induce acute stress among air craft crew members and general populations, for use in medical populations. We gathered continuous electrocardiography (ECG) data while medical students completed four different versions of the MATB-II of varying difficulties alongside questions probing medical knowledge, comprising a new task called medically-focused multitasking game (MFMG). After completing each version, participants responded to questionnaires to assess subjective states of stress (State Trait Anxiety Inventory for Adults) and cognitive workload (NASA Task Load Index). Responses to these questionnaires, physiological data from continuous ECG, and overall performance scores were combined to determine one version of MFMG that represented the highest level of elicited stress, and one that represented the lowest level of elicited stress. The results of this pilot study are promising, and have converged to reveal one high-stress and one low-stress version of MFMG, which will later be used to induce acute stress in biofeedback intervention studies among surgical residents and fellows. Beyond this specific application, MFMG can have broader applications in measuring acute stress induction and/or reduction among populations of healthcare practitioners.

INTRODUCTION

Acute stress is a well-documented problem for surgeons and surgical performance, especially novice surgeons (Arora, Sevdalis, Nestel, et al., 2010; Hassan et al., 2006; Healy & Tyrrell, 2011). Excessive levels of intraoperative stress are associated with impaired cognition (Wetzel et al., 2006) and psychomotor performance (Arora, Sevdalis, Aggarwal, et al., 2010). Surgery represents an acute care setting particularly susceptible to errors (Gruen, Jurkovich, McIntyre, Foy, & Maier, 2006), which may be exacerbated by excess stress. For these reasons, managing acute stress is imperative to optimize surgical performance and limit medical errors.

Biofeedback is one example of a quickly advancing technological tool that can mitigate adverse effects of acute stress (Kennedy & Parker, under review). Biofeedback training, often conveying the user’s heart rate in real time, enables the user to manage their physiological activity to improve health and performance (Ortiz-Vigon Uriarte, Garcia-Zapirain, & Garcia-Chimeno, 2015). As information is fed back to the user, voluntary and involuntary alterations in cognition, emotion, and behavior affect physiology, and the process continues iteratively (Schwartz, 2010). By monitoring physiological indicators of stress in real time, users could potentially learn to preemptively acknowledge and manage elevating stress levels.

Effective use of visual and auditory biofeedback among soldiers in simulation (Bouchard, Bernier, Boivin, Morin, & Robillard, 2012) offers convincing evidence for a similar benefit during acute stress simulation-based training for novice surgeons. Among surgical populations, biofeedback has also been shown to effectively reduce chronic stress levels (Lemaire, Wallace, Lewin, de Grood, & Schaefer, 2011). What remains to be seen is whether acute stress management can be accomplished using biofeedback in surgery.

Answering this question in the real-world setting has potential to introduce too much variability and confounding factors, minimizing the degree of control over the situation. The gap outlined represents one that first must be investigated in a setting free of excessive external influences, relying on a lab-based design. To reach these aims, a computer-based task must be investigated.

Since biofeedback in this capacity functions as an acute stress management tool, to evaluate its effectiveness we must first induce the stress we aim to mitigate. A variety of stress-inducing tasks evaluated through a meta-analysis of 208 lab-based experiments revealed that those containing uncontrollable and social-evaluative elements were associated with the largest increases in physiological stress (Dickerson & Kemeny, 2004). Given the uniqueness and complexity of the specific target population and its corresponding work domain, we must consider how researchers in the past have induced acute stress in other high-risk professions with stress levels approximating those seen in surgery.
Multi Attribute Task Battery (MATB-II), a well-validated and reliable tool, contains the key element of uncontrollability and was originally designed by NASA to induce stress among aircraft crew members and pilots (Comstock & Arnegard, 1992). Its applicability to other complex work domains with frequent exposures to acute stress is a key feature of MATB-II, due to its focus on simultaneous monitoring of multiple tasks at once. However, MATB-II is not specifically medically-relevant and does not address the crucial cognitive process of memory recall critical to medical decision-making.

To address these gaps, our lab adapted the MATB-II into a new task incorporating medical knowledge. This new tool, referred to as Medically-Focused Multitasking Game (MFMG), integrates three of the four MATB-II subtasks with timed medical questions. By combining MATB-II subtasks with medically-relevant questions, we have developed a new task that requires multitasking, ongoing monitoring and adjustment, and medical knowledge that is appropriate for the intended setting and population: surgical residents and fellows.

Prior to using MFMG in experimental settings, it is necessary to assess its capacity to appropriately induce stress, and to determine the impact of the medical questions on stress performance. MATB-II output represents performance on multiple subtasks, each of which can be manipulated to arrive at various versions of task difficulty levels. The degree of difficulty of the overall MFMG, which is largely determined by the degree of difficulty of MATB-II, has the potential to contribute to a variety of response patterns, some of which are undesired. Generating a level of difficulty that is too under- or too overwhelming could easily lead to boredom and/or disengagement, respectively. The task, therefore, needs to be stressful enough to induce the appropriate physiological and cognitive response, but not so stressful that disengagement ensues.

This study aims to address this process of realizing two independent versions of MFMG that represent appropriately low and high levels of stress in order to evaluate a stress management intervention in a population of surgical residents and fellows. We pilot tested versions of MATB-II in combination with medical questions (together comprising MFMG) among a group of medical students interested in the surgical domain. Analysis of physiological indicators of stress, subjective responses, and overall performance score determined the appropriate amount of stress for one low stress and one high stress version of MFMG.

We expect that subjective, objective, and performance results will align and thereby indicate which condition is considered the least stressful and which is considered the most stressful. This knowledge will inform the experimental design of future studies evaluating the most effective timing of presentation of components of an integrated biofeedback display with supplemental coping instructions. The display evaluated in future studies will have long-term implications for the appropriate integration of biofeedback as an acute stress management tool for novice surgeons during simulated stressful scenarios.

METHODS

Participants

A total of 10 medical students (5 females) were recruited to participate in this study. The participants included students across all years (3 first-years, 2 second-years, 2 third-years, 3 fourth-years), and 9 of the 10 participants were right-handed. The experiment was approved by and conducted in accordance with the guidelines of the Carilion Clinic Institutional Review Board, and subjects provided verbal consent before the study began.

Materials

Electrocardiography (ECG) signals were recorded using a 3-lead Mobile Cardio acquisition system (MindWare Technologies LTD, Gahanna, OH). This system acquired continuous heart rate (HR) and heart rate variability (HRV) data by affixing three lead-shielded electrode leads to disposable ECG electrodes transmitting voltage signals from the chest. Data was transferred via SD card to MindWare’s BioLab software and HRV application for analysis.

The MATB-II task was completed on a Dell desktop computer, while the medical questions were displayed and attended to simultaneously on an 8-inch Asus ZenPad 8.0.

Procedure

After going through the information sheet and providing verbal consent, participants were equipped with ECG electrodes and five minutes of baseline data were collected as the participant relaxed. Participants then viewed a 12-minute tutorial overviewing the task requirements of the following MATB-II subtasks: resource management, system monitoring, and tracking (Figure 1).

The main goal of the resource management task is to maintain 2500 units worth of fuel in the main tanks (A and B), despite their constantly depleting nature and occasionally broken pumps. The system monitoring task requires participants to detect and correct when the gauges deflect too far from the center, and when the lights are inappropriately on or off. To successfully perform the compensatory tracking task, participants must use a joystick to maintain the aircraft’s position (indicated by the circle) within the dotted square in the center. After summarizing the tasks and asking any questions, participants spent 20 minutes practicing the entire task, including all three MATB-II subtasks and the timed questions on the Asus tablet. Questions in this familiarization phase were not medically-oriented, but were comprised of random trivia knowledge with the purpose of acquainting participants with the appropriate gestures and multi-tasking requirements. Participants then filled out a brief demographic questionnaire with no identifying information and form Y-2 of the State Trait Anxiety Inventory for Adults (STAI) to assess the stable trait of overall stress levels.

The experimental phase consisted of completing four, 5-minute versions of the MFMG task. See Table 1 for details on the frequency of events contributing to the level of demand.
Medical questions represented a variety of specialties, ranging from neurosurgery to gerontology to psychiatry, for example. One medical student and one school of medicine faculty member reviewed the questions and determined them to be appropriate for gauging knowledge of medical residents and fellows, with potential difficulty foreseen in assessing medical students’ medical knowledge. Following each condition, participants responded to subjective measures of state stress (State Trait Anxiety Inventory for Adults, form Y-1; STAI) and cognitive/mental workload (NASA Task Load Index; NASA-TLX). At the completion of the fourth condition, participants weighed each workload dimension in the NASA-TLX.

**Analysis**

Total scores for the STAI and NASA-TLX were calculated using Excel and the means were compared using repeated measures ANOVAs calculated in JMP Pro 12 by SAS Institute, Inc. (Cary, NC). These data were used to provide a more subjective indicator of the most and least difficult versions of the overall task.

HR and HRV analysis were conducted using MindWare’s BioLab acquisition and HRV analysis modules. Calculations of time-domain features of HRV such as standard deviation from normal-to-normal (SDNN), root mean square of the successive differences (RMSSD), and percentage of consecutive N-N peaks differing by 50 milliseconds (pNN50) provide us with measures of overall variability and beat-to-beat variability. Five minutes of data were selected from the 20-minute practice period, from the time range of 10-15 minutes, to derive vanilla baseline values of each of the metrics analyzed. SDNN, RMSSD, and pNN50 values were then normalized for every individual based on their values during the vanilla baseline period, and aggregated across participants. Results from repeated measures ANOVAs using JMP Pro 12 provide a more objective ranking of stress.

Performance analysis using Microsoft Excel and JMP Pro 12 provides an overall performance score by assigning equal weight to the three MATB-II subtasks and combining them. Performance on the resource management subtask was represented by calculating the average deviation from 2500 units in tanks A and B. To represent performance on the system monitoring task, we calculated the proportion of correct responses to deflections. Performance on the tracking task was represented by the root-mean-square (RMS) error. Finally, performance on the medical assessment questions was evaluated by the proportion of correct responses. Following these analyses, performance on each subtask was normalized according to overall performance on that individual’s subtask in the practice period, creating a z-score, with the exception of the medical assessment subtask. This exception was made because questions administered during the practice period were intentionally non-medical and represented random knowledge. The nature of the practice questions was not comparable to that of the medical questions, or representative of the pre-acquired knowledge this subject pool was expected to have. The z-scores from the MATB-II subtasks and a raw score representative of medical assessment performance were then summed to create one value representing overall performance, with each subtask receiving equal weight. All performance scores were averaged across participants within each version of MATB-II, and those means were compared using repeated measures ANOVAs in JMP Pro 12.

**RESULTS**

**Subjective Indicators**

Subjective states of stress and cognitive workload were analyzed by conducting repeated measures ANOVAs on scores on the STAI and NASA-TLX, respectively. While results from the STAI show no significant differences between conditions, the results from the NASA-TLX do reveal significant differences (Figure 2). Specifically, comparisons of means using paired t-tests indicated that there was a significant difference between the mean perceived cognitive workload in MATB-A (M=74.33, SD=14.13) and both MATB-C (M=61.21, SD=16.09); t(9)=3.99, p=0.0016 and MATB-D (M=63.23, SD=13.15); t(9)=2.15, p=0.0302, supporting a significantly higher perceived cognitive workload elicited by the most difficult condition compared to the two easiest conditions. There was also a significant difference between the mean perceived cognitive workload in MATB-B (M=72.15, SD=10.67) and MATB-D (M=63.23, SD=13.15); t(9)=2.01, p=0.0379, supporting a significantly higher workload elicited by the condition with the second highest difficulty compared to that with the lowest difficulty. No other significant differences were found between means.

**Physiological Indicators**

Data from continuous ECG collection were analyzed by individual HRV components and overall HR and subjected to repeated measures ANOVAs. Analysis of the mean HR across conditions, normalized to each individual’s baseline, reveals no significant differences and no general trends (Figure 3).

Comparisons of means via paired t-tests revealed that there was a significant difference between RMSSD values in MATB-A (M=0.02, SD=0.18) and MATB-C (M=0.14, SD=0.24); t(8)=2.79, p=0.0179. This difference supports a significantly lower value for this physiological indicator of stress, RMSSD, corresponding to a higher state of stress, in the most difficult version of the task compared to the second easiest version. Overall, values across all four conditions follow the expected linear relationship of increasing HRV values with decreasing levels of task difficulty. The lowest HRV values are seen in the most difficult version of the task and the highest HRV values are seen in the easiest version of the task. Comparison of means among pNN50 results across conditions reveals no significant differences (Figure 4).

Paired t-tests to evaluate the differences between SDNN values indicate significant differences between MATB-A (M=0.11, SD=0.12) and all other conditions [(MATB-B: M=0.02, SD=0.20; t(8)=2.28, p=0.0261); (MATB-C: M=0.00, SD=0.20; t(8)=2.18, p=0.0303); (MATB-D: M=0.09, SD=0.23; t(8)=2.35, p=0.0233)]. These results support that the
value for this physiological indicator of stress, SDNN, is significantly lower, corresponding to a higher state of stress, in the most difficult version of the task compared to all other versions (Figure 5).

**Performance Indicators**

An overall performance score for each version of MFMG was derived (see Analysis for details). Overall performance was analyzed by comparing the means across conditions through repeated measures ANOVA. These analyses revealed no significant differences (Figure 6).

**DISCUSSION**

This pilot study demonstrated a convergence of data from various sources, including subjective reports, performance measures, and physiological recordings. Results suggest that there are two disparate versions of the same task which can induce differential levels of physiological and psychological stress.

Although the two subjective stress measures collected did not yield consistent results, we can say with some confidence that the responses via the NASA-TLX are more representative of the cognitive workload we intended to measure, rather than those from the STAI. While the STAI has historically been referred to as a sensitive and reliable measure of anxiety, it is not without its barriers (Marteau & Bekker, 1992). The confidence in our results extends to the frequent and reliable use of the NASA-TLX among medical populations (Alaraj, Tobin, & Birk, 2013; Alkahtani, Aziz, Ahmad, & Darmoul, 2015; Andersen, Klein, Gögenur, & Rosenberg, 2012; Crewther et al., 2016; Effken, Loeb, Kang, & Lin, 2008; Mazur et al., 2014; Park et al., 2017; Wadhera et al., 2010; Yurko, Scerbo, Prabhu, Acker, & Stefanidis, 2010; Zheng et al., 2012), as well as the development and validation of the SURG-TLX (Berg et al., 2015; Hallbeck, Lowndes, & Bingener, 2013; Lowndes, Bingener-Cassey, & Hallbeck, 2014; Roy et al., 2015; Weigl et al., 2016; Wilson et al., 2011), an adaptation specific to surgeons. Our future work in this area will address these considerations by administering the short-form version of STAI developed by Marteau and Bekker in 1992, as well using the SURG-TLX scale.

Previous work has called into question the validity of using HR as a metric for performance or stress in surgical populations (Goldman, McDonough, & Roemond, 1972; Payne & Rick, 1986). Additional work has specifically suggested an enhanced sensitivity among HRV measures compared to HR measures to detect acute stress and mental strain among surgeons (Böhm, Rötting, Schwenk, Grebe, & Mansmann, 2001).

Results from the comparison of means for pNN50 did not reach significance at \( \alpha = 0.05 \), but do follow a linear trend. On the other hand, the trends for SDNN and RMSSD are in the expected direction and comparisons of means do reveal significant differences between conditions. The primary limitation affecting lack of significance in physiological data is likely the small sample size recruited (n = 10), but results are still promising.

Data representing overall performance showed a trend similar to that of cognitive workload measured by the NASA-TLX, in which MATB-A and MATB-B versions of the task showed poorer performance and were closer in value to one another, while MATB-C and MATB-D versions of the task showed better performance and were closer in value to one another. The comparisons of means between performance scores across conditions revealed that the differences were not significant at \( \alpha = 0.05 \), however. This lack of significant findings may point to the multiple different ways to capture optimal performance on each subtask, and the nuances among the interactions between subtasks.

While performance on the system monitoring subtask was represented by the proportion of correct responses to observed deflections, an alternative metric could have been reaction time. We intentionally chose not to use reaction time due to the nature of the physical set-up of the experiment. With the medical assessment questions displayed on a separate tablet next to the desktop computer, and the remaining three subtasks displayed in one interface and on one computer screen, there was a tendency for participants to focus more on the MATB-II interface when the MATB-II had a higher frequency of changing events, and more on the medical assessment questions when the MATB-II had a lower frequency of changing events. This would result in MATB-A having either very quick reaction times while attending to frequent deflections in system monitoring and many missed medical questions, or many missed deflections in the system monitoring task while participants are attending to the medical questions. On the other hand, MATB-D with a very low frequency of events might encourage more attention to the medical questions on the tablet and a slower reaction time in terms of deflections in the system monitoring subtask. When reaction time was included as the representative performance metric in the systems monitoring subtask, replacing the proportion of correct responses, the difference in overall performance scores across conditions was negligible (data not shown).

Another concern that could have affected performance on the overall MFMG task was the high potential that scores on the medical assessment subtask reflected the proportion of correctly guessed answers rather than the proportion of known answers. As medical students in the midst of training, a majority of participants admitted to guessing answers. To account for this potential confound, and the potential that some sets of questions may have been easier than others, an overall performance score was calculated excluding the contribution from medical assessment. As a result, overall performance in each condition decreased by roughly the same amount (data not shown), indicating that it was in fact the disparate versions of the MATB-II that contributed almost exclusively to the disparate difficulty of the MFMG task.

The potential for knowledge advancement in the realm of investigations into acute stress interventions in healthcare are enormous as a result of this work. By generating and identifying appropriately stressful versions of what is essentially a “medical MATB-II,” the opportunity to evaluate acute stress using well-established and domain-
relevant tasks in a tightly controlled setting can be realized. Investigations into basic processes associated with acute stress and acute stress recovery can be systematically accomplished using this task, which will set the stage for larger-scale investigations into similar processes in an applied setting in the future.

The activation of existing cognitive processes crucial for successful surgical performance, such as multi-tasking, psychomotor control, dynamic resource management, decision-making, uncontrollability, and attention allocation can be accomplished successfully through the use of MATB-II. But to the best of our knowledge, MATB-II has never been used in empirical research to evaluate the stress response in a healthcare setting or among a population of healthcare practitioners. By complementing this task with the added component of memory recall and making it medically-relevant with the addition of medical assessment questions, the total task (MFMG) can more appropriately address additional facets of surgeons’ cognitive processes than either task alone.

Thus, the value in this work has direct real-world applications. We now have an appropriate and domain-relevant, stress-inducing, lab-based task to evaluate the effectiveness of a stress intervention paradigm in healthcare, under tightly controlled settings. The applicability of a task as valid and reliable as MATB-II into the healthcare setting will extend the possibilities and research avenues within a population of healthcare providers.

ACKNOWLEDGMENTS

Agency for Healthcare Research and Quality R18HS023465-02 (PI- Parker)

REFERENCES


Figure 1. MATB-II interface. The MF MG task incorporated the following subtasks seen in this interface: system monitoring (top left), tracking (top center), and resource management (bottom center). The communications subtask was not included (bottom left). Medical questions were displayed on a separate screen.

Figure 3. The difference between normalized mean HR values across conditions was not significant, and there is no observed trend. This is especially apparent when considering the range of values across conditions: 0.007 units.

Figure 5. The normalized mean SDNN value in the condition with the highest frequency of events (MATB-A) was significantly lower than that value in all remaining conditions at $\alpha = 0.05$.

Figure 2. The mean perceived level of cognitive workload across participants from the condition with the highest frequency of events (MATB-A) was significantly higher than the conditions representing the lowest frequency of events (MATB-D) and the second lowest frequency of events (MATB-C) at $\alpha = 0.05$. Furthermore, the condition with the second highest frequency of events (MATB-B) had a significantly higher mean than MATB-D ($p<0.05$).

Figure 4. The difference between normalized mean pNN50 values across conditions was not significant, but the observed trend aligns with our expectations. The normalized mean RMSSD value in the most difficult condition (MATB-A) was significantly lower than that in the second easiest version of the task (MATB-C) at $\alpha = 0.05$.

Figure 6. Overall performance scores, grouped by condition. There are no statistically significant differences between the means across conditions, but the expected trend is still supported. The conditions with the highest frequency of changing events (MATB-A and MATB-B) also have the lowest overall performance score across participants, while the conditions with the lowest frequency of changing events (MATB-C and MATB-D) both have the highest overall performance score across participants.
Table 1. Characteristics for each subtask of MFMG. The frequency of events within each MATB-II subtask differs depending on the version of MFMG being played, while the medical assessment questions are always randomized. The medical assessment portion of MFMG should not differ in difficulty or affect the overall difficulty across versions.

<table>
<thead>
<tr>
<th>Condition</th>
<th>Tracking</th>
<th>System monitoring</th>
<th>Resource management</th>
<th>Medical questions</th>
</tr>
</thead>
<tbody>
<tr>
<td>MATB-A</td>
<td>Medium default</td>
<td>30 deflections/min</td>
<td>1-2 pumps fail for 15 seconds every minute</td>
<td>Randomized order</td>
</tr>
<tr>
<td>MATB-B</td>
<td>Medium default</td>
<td>20 deflections/min</td>
<td>1-2 pumps fail for 5 seconds every minute</td>
<td>Randomized order</td>
</tr>
<tr>
<td>MATB-C</td>
<td>Easy default</td>
<td>10 deflections/min</td>
<td>1 pump fails for 15 seconds every minute</td>
<td>Randomized order</td>
</tr>
<tr>
<td>MATB-D</td>
<td>Easy default</td>
<td>2 deflections/min</td>
<td>1 pump fails for 5 seconds every minute</td>
<td>Randomized order</td>
</tr>
</tbody>
</table>