COMPUTATIONAL AND STATISTICAL MODELING OF THE VIRTUAL REALITY STROOP TASK

Justin M. Asbee

Dissertation Prepared for the Degree of

DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

May 2022

APPROVED:

Thomas Parsons, Committee Co-Chair
Kimberly Kelly, Committee Co-Chair
Heidemarie Blumenthal, Committee Member
Timothy (Fred) McMahan, Committee Member
Anthony Ryals, Committee Member
Donald Dougherty, Chair of the Department of Psychology
Tamara L. Brown, Executive Dean of the College of Liberal Arts and Social Sciences
Victor Prybutok, Dean of the Toulouse Graduate School

Asbee, Justin M. *Computational and Statistical Modeling of the Virtual Reality Stroop Task*. Doctor of Philosophy (Behavioral Science), May 2022, 73 pp., 22 tables, 17 figures, references, 94 titles.

The purpose of this research was two-fold: (1) further validate the Virtual Reality Stroop Task HMMWV [VRST; Stroop stimuli embedded within a virtual high mobility multipurpose wheeled vehicle] via a comparison of the 3-dimensional VRST factor structure to that of a 2-dimensional computerized version of the Stroop task; and (2) model the performance of machine learning [ML] classifiers and hyper-parameters for an adaptive version of the VRST. Both the 3-D VRST and 2-D computerized Stroop tasks produced two-factor solutions: an accuracy factor and a reaction time factor. The factors had low correlations suggesting participants may be focusing on either responding to stimuli accurately or swiftly. In future studies researchers may want to consider throughput, a measure of correct responses per unit of time. The assessment of naive Bayes (NB), k-nearest neighbors (kNN), and support vector machines (SVM) machine learning classifiers found that SVM classifiers tended to have the highest accuracies and greatest areas under the curve when classifying users as high or low performers. NB also performed well but kNN algorithms did not. As such, SVM and NB may be the best candidates for creation of an adaptive version of the VRST.

TABLE OF CONTENTS

Page

LIST OF TABLES

Page

LIST OF FIGURES

Page

CHAPTER 1

INTRODUCTION

1.1 Stroop Task

A common tests of executive functioning is the Stroop task (Scarpina and Tagini 2017). Psychologists have broadly defined executive functioning as the ability to control complex cognition and engage in thoughts and behaviors related to current goals, while ignoring irrelevant stimuli (McCabe et al., 2010). Executive functioning consists of multiple aspects of cognitive control abilities (e.g., cognitive workload, attention, planning, goal orientation, and inhibition; (McCabe et al., 2010). The Stroop task is believed to mainly assess inhibition, to produce a correct response participants need to rely on controlled processing to inhibit automatic responses (Heidlmayr et al., 2020).

1.2 Theories Explaining the Stroop

When performing the Stroop task several conditions exist. Participants are often asked to name patches of colors, read color words, and name colors (MacLeod, 1991). The Stroop effect (i.e., an increase in response time due to an interference effect) occurs when participants are asked to name the color of the stimuli when the semantic meaning of the color word does not match the font of the word (e.g., the word blue presented in red font, e.g., blue; see Figure 1.1).

Figure 1.1

Example Stimuli from the Automated Neuropsychological Assessment Metrics (ANAM) Stroop Task

Macleod (1991) reviews several theories used to explain the Stroop effect such as perceptual encoding. According to this theory the encoding of information such as a neutral control or congruent stimuli (i.e., where the ink color and word color match) occurs more quickly than the processing of incongruent stimuli because the information provided by the ink color and word are not compatible. Another explanation is the relative speed of processing model. According to the relative speed of processing model both naming of the color and reading of the word are processed at the same time. However, to produce a response a threshold of activation must be reached before a response is produced. This model assumes that interference is produced because word reading can be processed faster than color naming. There is a resultant increase in time required for a correct response.

While there are other models aimed at explaining the Stroop effect, the predominant view is that the Stroop effect is likely due to controlled and automatic processing (Lifshitz et al., 2013). According to dual process theory, automatic and controlled processing are two distinct systems. These two systems are used for decision making and stimulus responses (Pennycook, 2017). Lifshitz and colleagues (2013) state automatic processes are rapid, automatic, and non-conscious, controlled processes by contrast are effortful, slow, and deliberate. The Stroop effect is thought to involve controlled inhibition of automatic (e.g., word reading) responses, which requires greater cognitive resources and leads to increased response time (Lifshitz et al., 2013). However, scores on the Stroop task have been found to be correlated with and may in part rely other cognitive functions such as attention, processing speed, and working memory (Periáñez et al., 2020)

1.3 Low-Dimensional Computer-Automated Stroop Presentations

Many Stroop tasks are considered low dimensional assessments. Low dimensional

assessments are tasks or measures that often do not allow for interactivity and utilize static stimuli which may not reflect real-world situations (Parsons & Duffield, 2020). Many of the earliest examples of low dimensional assessments are paper-and-pencil measures using simple stimuli (Parsons & Duffield, 2020). Low dimensional assessments typically involve little contextual information, (i.e., lacking much of the environmental cues readily available in real-world settings). The addition of environmental and contextual cues may be important for generalization to situations and environments outside of the lab setting (Schilbach, 2015).

Computerized assessments represent a step beyond simple paper-and-pencil measures. Computer-based assessments allow for some degree of automation. Automation can range from presentation of directions or stimuli to scoring and evaluation. Computerized assessments may allow for precise measurements of behaviors such as reaction time (Rabin et al., 2014). Computerized assessments have improved upon traditional measures; however, many can still be considered low dimensional assessments (Parsons & Duffield, 2020). Computerized assessments are typically designed to mirror the traditional psychological measures they were based on, leading many to have the same issues as traditional assessments (Kessels, 2019). Unfortunately, many lower dimensional assessments are still considered the gold standard for measurement of psychological constructs. This is often due to the large amounts of norming data that have been collected over time (Woodhouse et al., 2013). Lower dimensional measures have advanced our knowledge and understanding within the field of psychology, but these measures are not without their limitations (Pan & Hamilton, 2018).

The Automated Neuropsychological Assessment Metrics (ANAM) version of the Stroop task (Vista Life Sciences) is an example of a computer automated Stroop task (Reeves et al., 2007). Participants perform three Stroop conditions within the ANAM Stroop, the primary

outcome measures are speed and accuracy (Reeves et al., 2007). Computerized versions of the Stroop task often involve single stimulus presentations of Stroop stimuli, which may minimize interference from surrounding Stroop stimuli (Periáñez et al., 2021). Further, computerized Stroop tasks may offer reaction times for individual Stroop stimuli. Participant performance for congruent (i.e., word font and written word match) or incongruent (i.e., word font and written word do not match) stimuli may also be assessed independently with single-item presentations (Brunetti et al., 2021).

Many low dimensional psychological assessments emphasize experimental control which may lead to a decreased ability of assessments to reflect real-world outcomes (Parsons, 2015). For many of these lower dimensional assessments application outside of the lab may not have been their intended purpose, but many of these assessments have been adopted for such use (Baumeister, 2016). Often this approach leads to issues involving ecological validity where the ability of low dimensional tasks to predict real-world behavior is poor (Pan & Hamilton, 2018). Many low dimensional assessments are designed to optimize experimental control while sacrificing ecological validity. Parsons (2015) states that ecological validity has two major components: veridicality and verisimilitude. Veridicality infers that construct-driven measures should predict functioning in daily life (e.g., measures of memory should be related ability to remember items while shopping). Verisimilitude is an indication of the degree to which the psychological measures and testing conditions resemble the users' daily activities (Parsons, 2015).

1.4 Cognitive and Affective Processing

Additionally, many low dimensional psychological assessments focus on abstract cognitive tasks and emotional/affective processing is often not assessed. Zelazo (2015) argues

that both cognitive and affective processing are important in many real-world situations. Assessments focused on cognitive abilities tend to be more related to abstract processes such as planning, inhibition, or working memory (Nejati et al., 2018). A potential reason lower dimensional assessments may not tap into emotional processing is because of the lack of contextual/environmental cues. Cognitive processing involved during low-dimensional Stroop tasks is believed to occur in parts of the anterior cingulate cortex (ACC; Cieslik et al., 2015). Ruff and colleagues (2001) found that the dorsal aspect of the ACC was most activated during conflict (i.e., incongruent conditions) within a lower dimensional Stroop task. The Stroop task is also believed to involve other areas connected to the dorsal ACC, mainly the dorsolateral prefrontal cortex (DLPFC; Cieslik et al., 2015). However, many other brain areas are involved in executive functioning. For example, working memory is unlikely to rely solely on the DLPFC (Diamond & Levine, 2018), interactions between brain areas such as the medial prefrontal cortex (mPFC) and hippocampus may be critical for components of cognitive processing such as working memory (Jin & Maren, 2015).

Areas of the ACC play an important role in both abstract and emotional processing (Shenhav et al., 2016). may not be fully assessed by lower dimensions measures Tasks that are affective in nature typically involve emotion, motivation, or immediate vs late delayed gratification (Zelazo, 2015). Ventral aspects of the ACC are important for decision making regarding risk and reward (Cai & Padoa-Schioppa, 2012). Ventral aspects of the ACC may also play a role in effort and motivation potentially determining how much attention is given to cognitive assessments (Shenhav et al., 2016). Ventral aspects of the ACC connect to emotional areas of the brain such as the limbic area and to some frontal regions such as the orbitofrontal

cortex (OFC; Feroz et al., 2019). The OFC and the limbic area play key roles in emotional and affective processing (Zelazo, 2015).

Many of these brain areas may also be important for emotional responses (Schweizer et al., 2013). Increases in cognitive or affective load (i.e., the degree to which these systems are utilized) may lead to strains in both systems and possibly lead to reductions in inhibitory/control abilities (Plass & Kalyuga, 2019). The modal model of emotions states that the evaluations of stimuli can lead to changes in cognition and physiology based on whether these stimuli are attended to or not (Gross, 2015). A person's active goals and the ability or resources available to dynamically attend to these goals likely influence which stimuli are attend to (Gross, 2015). Cognitive load has also been found to be correlated with multiple measures of psychophysiological arousal including electrodermal response (Romine et al., 2020), heart rate (Johannessen et al., 2020), and respiration rate (Barua et al., 2020). Research on the Stroop task, typically considered a cognitive task, found changes in psychophysiological arousal occur during the assessment, measured via pupil diameter (Laeng et al., 2011).

1.5 Virtual Reality (VR) Assessments

Leveraging newer technologies for psychological assessment may provide advantages over previous methods. A range of technologies which may provide enhancements to current methods include manipulation of measurement techniques, computational modeling and simulation, and virtual reality (VR; Parsons & Duffield, 2019). Typically, VR assessments are a move beyond computerized assessments and may enable researchers to create and present users with high dimensional, interactive stimuli (Pan & Hamilton, 2018). These higher dimensional, interactive stimuli may increase ecological validity of the assessment. High dimensional tools such VR assessments, may help with some of the ecological validity issues typically associated

with 2D computerized assessments. For example, VR may offer enhanced representation of real-world environments, engaging background narratives, enhanced stimulus presentations, and diverse (while reliable) ways of presenting stimuli (Gerjets et al., 2014; Parsons 2015). Most low dimensional tasks generally have simple stimulus presentations (Parsons & Duffield, 2020). Stroop task stimuli are traditionally presented concurrently on a card or sheet of paper. Moving to computerized assessments can be an improvement over traditional assessment, however as mentioned before many computerized assessments do not address many of the issues associated with traditional low dimensional assessments (Kessels, 2019).

High dimensional VR assessments present users with more environmental information and do not simply provide users with stimuli devoid of context (e.g., a sheet of stimuli or stimuli centered within a blank screen; Parsons et al., 2013). For example, the Virtual Reality Stroop Task (VRST; HMMWV version) has participants ride in a simulated high mobility multipurpose wheeled vehicle (HMMWV) along a desert road modeled after middle eastern environments (Parsons et al., 2013). Higher dimensional assessments may lead to increased levels of immersion due to increased levels of detail provided by the stimuli and possibly leading to arousal (Diemer et al., 2015). Moreover, Parsons and associates (2011) found that when participants were in a more immersive virtual environment (i.e., use of VR equipment such as a head mounted display), they had increased levels of arousal compared to a less virtual immersive environment, where they interacted with the environment using a standard laptop display.

Neguţ and colleagues (2015) conducted a meta-analysis to examine convergent validity of assessments utilizing VR with lower dimensional assessments. The results suggested moderate convergent validity. The researchers stated that, often significant differences exist between lower dimensional tests and measures utilizing high dimensional VR. These differences may be due to

many VR assessments being designed to have greater ecological validity (Neguţ et al., 2015; Parsons et al., 2017).

1.6 The Virtual Reality Stroop Task (VRST)

The VRST (HMMWV version; Figure 1.2) provides detailed environmental information to participants. The VRST manipulates various aspects of the environment (i.e., number of arousing stimuli) and stimulus presentation complexity (i.e., congruency of Stroop stimuli and location of stimuli) while participants perform the Stroop task (Parsons & Courtney, 2018). The VRST includes four conditions: word reading, color naming, simple interference, and complex interference. The word reading, color naming and simple interference conditions include Stroop stimuli that are like other Stroop tasks. Within the VRST stimuli are presented in the middle of the user's visual field except for the complex interference condition where stimuli are presented at varying locations in the user's visual field (Parsons et al., 2013).

Figure 1.2

Screen Captures from the VRST, Images Show the Complex Interference Condition

The VRST incorporates some aspects lower dimensional Stroop tasks but also includes additional information via the environment. While there are differences between the Stroop

modalities, previous research has observed the overall Stroop effect in conditions where giving the correct response required more cognitive resources and in general involved an increase in response times (Parsons & Barnett, 2019, 2018). Previous research has found correlations between scores from low dimensional Stroop tasks and the VRST, indicating construct validity (Armstrong et al., 2013; Parsons et al., 2013).

Participants experience all four Stroop conditions from the VRST in both safe and ambush zones. The VRST is believed to influence affective processing, particularly within ambush zones. The ambush zones are marked by a high number of potentially arousing stimuli (e.g., explosions and gunfire) compared to safe zones which presents a minimal number of external distractors. The ambush zones have been found to increase several measures of autonomic arousal including heart rate, respiration rate, and skin conductance level, suggesting changes in affective and cognitive load (Parsons & Courtney, 2018). As previously suggested users may have to direct attention away from the environment and potentially arousing stimuli to increase task performance (Parsons & Courtney, 2018).

Overall performance on the VRST may be impacted by levels of arousal in either a positive or negative way. According to Rozenek and colleagues (Rozenek et al., 2019) typically cognitive performance for moderately challenging tasks is best when under moderate levels of arousal. When arousal is too high cognitive performance tends to suffer as emotions such as fear or rage may produce restricted awareness and disorganized behavior to occur (Rozenek et al., 2019). Performance may also decline with low levels of arousal, when arousal is insufficient drowsiness and disengagement may occur (Rozenek et al., 2019; Wekselblatt & Niell, 2015). The inverted-U shape of performance (i.e., lowered performance under low or high arousal) has been called the Yerkes-Dodson Law (Chaby et al., 2015). The Yerkes-Dodson Law is based on a

series of experiments where moderate levels of arousal were associated with the greatest levels of task performance (Chaby et al., 2015).

Previous research modeled optimal arousal and performance within the VRST (Wu et al., 2010). Researchers found that reaction times followed a pattern like the Yerkes-Dodson Law, where optimal performance is found at moderate levels of arousal (Wu et al., 2010; Parsons & Reinebold, 2011). Wu and colleagues (2010) found that psychophysiological measures collected during the VRST could be used to evaluate task difficulty. In a series of studies Wu and colleagues used machine learning (ML) algorithms to examine arousal data from the VRST (Wu et al., 2010; Wu & Parsons, 2011a; Wu & Parsons, 2011b; Wu & Parsons, 2012; Wu et al., 2013). Wu and associates (2010) initially examined the feasibility of identifying optimal arousal in participants. The researchers found that there was a high degree of individual variability in arousal, using a classifier based on group level metrics for arousal likely led to decreased classifier performance. The next several studies attempted to reduce the amount of subject-specific data needed to accurately classify arousal. The researchers examined transfer learning, (Wu & Parsons, 2011a) active class selection, (Wu & Parsons, 2011b), and a combination of transfer learning and active class selection (Wu & Parsons, 2012; Wu et al., 2013).

1.7 Adaptive Assessments and Flow

Parsons and Reinebold (2012) created a framework for creating adaptive high dimensional assessment. Features of the assessment (e.g., number of stimuli or task difficulty) can be manipulated to allow users to potentially stay within a level of optimal performance (Rodríguez-Ardura & Meseguer-Artola, 2016). Parsons and Reinebold (2012) used transfer learning, active class selection, and a combination of the two based on the research by Wu and colleagues (2010; 2011a; 2011b) to improve classifications of participant's affective states. Data

from the assessment was used to evaluate task difficulty based upon participant performance and psychophysiological measures. These metrics were then used to allow for changes within the VR assessment.

The flow model may be helpful for predicting performance within VR assessments such as the VRST. The flow model suggests that trade-offs between environmental excesses and decrements (i.e., variation in cognitive and affective load) can lead to changes in task performance associated with experiences of flow (i.e., a state of immersion or absorption in the current task) (Nakamura & Csikszentmihalyi, 2014). Increased experiences of flow may lead to optimal levels of performance which may be associated multiple factors such as when task difficulty is consistent with the participant's skill level (Nakamura & Csikszentmihalyi, 2014). Other factors, which improve the likelihood of entering flow states other than the challenge-skill balance include unambiguous feedback and having a sense of control (Gerjets et al., 2014; Rodríguez-Ardura & Meseguer-Artola, 2016). Increases in experiences of flow have been associated with increased performance on a wide variety of tasks such as athletics, musicianship, and learning (Chirico et al., 2015; Norsworthy et al., 2017; Rodríguez-Ardura & Meseguer-Artola, 2016).

Adaptive assessments can be used to manipulate many of the factors associated with flow (Parsons & Reinebold, 2012). Well-designed VR assessments may be particularly suited for placing people in flow states because they can manipulate many aspects of the assessment associated with flow (McMahan & Parsons, 2020). The VRST for example, could increase or reduce the number of arousing stimuli presented in the environment, potentially leading to an optimal level of task difficulty for the user.

The classification algorithm (i.e., statistical procedure for assessing participant

performance) is an important aspect of adaptive VR assessments. Selecting and evaluating various ML algorithms is essential because some algorithms perform better in certain situations compared to others (Parsons et al., 2022). Classifying events related to participant reactions is necessary for adaptive VR assessments and more research evaluating ML algorithms would be beneficial (Parsons et al., 2022). Further, classification techniques may have utility for screening and assessing a variety of conditions and disorders such as autism spectrum disorder, traumatic brain injury, and post-traumatic stress disorder (Galatzer-Levy et al., 2017; Mitra et al., 2016; Omar et al., 2019). Within the clinical field classifiers may be able to be used as screeners to determine if users may need to pursue further treatment. For example, classifiers have been used to automated detection of traumatic brain injury (Mitra et al., 2016). Additionally, classification algorithms may potentially reduce the costs typically associated with autism spectrum disorder diagnosis as classifiers have had some success as screening tools for autism spectrum disorder (Omar et al., 2019). Additionally, Badesa and colleagues (2014) suggested that ML classifiers could be used to for adaptive therapy applications as well.

Adaptations within the VE occur in a variety of ways including: the addition and removal of stimuli, increasing and decreasing the number of distractions in the VE, or providing guidance to the participant (Drey et al., 2020; Zahabi & Abdul Razak, 2020). Parsons and colleagues 2022) examined EEG signals in the context of game related events. They assessed the performance of SVM, NB, and kNN, they found that NB produced the best classification for negative game-based events and kNN was best for classifying beta bands from EEG data (Parsons et al., 2022). McMahan and colleagues (2021) evaluated three classification techniques for building an adaptive Virtual Reality Stroop Task within a classroom setting. The researchers examined support vector machines (SVM), Naïve Bayes (NB), and k-nearest neighbors (kNN).

They found that SVM outperformed both NB and kNN classification algorithms when using a 10-fold cross validation (McMahan et al., 2021).

The following sections describe and explore the use of several of these technologies. The first study provides additional validation of the VRST, a 3-dimensional Stroop task, by comparing its factor structure with the factor structure of a 2-dimensional Stroop task the ANAM Stroop task. The second study examines several ML algorithms and hyper-parameters that can potentially enhance the classification of participant performance within the VRST for an adaptive version.

CHAPTER 2

FACTOR ANALYSIS OF THE VRST AND ANAM STROOP TASK

2.1 Purpose of Factor Analysis

While there is a growing body of literature on the VRST and its validity, the factor structure has yet to be explored and would provide a more detailed understanding of the VRST. Therefore, the current work performed an exploratory factor analysis of the VRST. Additionally, the factor structure of the VRST is compared to a computerized (ANAM) Stroop task.

2.2 Methods

2.2.1 Participants

The current study conducted an examination of Stroop tasks performed by undergraduate students (*N* = 115; *M* age = 20.39, *SD* = 3.56; 56.52% women) from a large university in the southwestern United States. The participants primarily identified as non-Hispanic white (48.48%; Table 2.1).

Table 2.1

*Demographics (*N = *115)*

Characteristic	Mean (SD)	%
Age (in years)	20.39 (3.56)	
Men		43.48
Women		56.52
Black		18.26
Asian		7.83
Hispanic		21.74
Non-Hispanic white		48.70
Other/Multiracial		1.74

Participants with scores on variables greater than two standard deviations from the mean had

those variable scores removed from the analyses (91 participants had data for all VRST conditions, 97 had data for safe zones, 102 had data for all ambush zones, and 98 had data for the ANAM Stroop). Participants completed a battery of assessments in addition to the ANAM and VRST.

2.2.2 Materials

2.2.2.1 ANAM (Automated Neuropsychological Assessment Metrics)

The ANAM Stroop is a computer-automated version of the Stroop task. This version of the Stroop task has three conditions word reading, color naming, and simple interference. In each of these conditions, participants are presented with 50 items. Participants are asked to respond to stimuli both verbally and by pressing a button. The word reading condition shows participants the words red, green, or blue, which are in all caps and white font. Participants are asked to respond to the color of the word. When performing the color naming condition participants respond to four capital X's in either red, green, or blue font. Participants are instructed to respond to the color of the font of the X's. Lastly, in the simple interference condition, participants are presented with color words (either red, green, or blue) which are in red, green, or blue fonts and are asked to respond to the color of the font rather than reading the word. Scores from the ANAM Stroop are collected automatically.

During the ANAM Stroop, participants used a standard desktop computer with a keyboard and mouse as interfaces. Participants are presented on a standard computer monitor with a black screen. Instructions are presented in white font. Before beginning, participants read instructions on the computer screen and are prompted to verbally respond by reading each word aloud as they press a color-coded button on their keyboard. Participants are asked to respond to each stimulus as quickly as possible without making mistakes. Each stimulus is presented one at

a time in the middle of the screen. Once the participant responds to an item the next item is displayed. Each condition consists of several practice items before informing the participant that they will be tested for speed and accuracy, and then testing the participant on 50 Stroop items.

2.2.2.2 VRST (Virtual Reality Stroop Task)

The high mobility multipurpose wheeled vehicle (HMMWV) version of the VRST (Figure 1.2) is a 3D presentation of the Stroop task. When performing the VRST participants ride in a simulated HMMWV along a desert road modeled after middle eastern environments. Participants are not required to drive the vehicle and the simulated HMMWV travels along a pre-determined path. The VRST consists of four Stroop conditions: word reading, color naming, simple interference, and complex interference. However, complex interference was not included in the current study to make results more comparable to the ANAM Stroop. Of the Participants first respond to color words, then colored X's, and then words in various font colors. The complex interference Stroop condition is like the simple interference condition with an additional degree of difficulty. Within the complex interference condition instead of the stimuli appearing in the middle of the windshield they appear in various locations of the windshield. Also, participants experience safe zones and ambush zones which are alternating and counterbalanced. Potentially arousing stimuli outside of the HMMWV are not included as it travels through the safe zones. In ambush zones participants experience arousing stimuli such as simulated gunfire, explosions, and shouting. Participants experience each of the Stroop conditions in both safe and ambush zones (eight conditions in total).

The VRST may be displayed using either a standard computer monitor or can be displayed through a head mounted display (HMD). In the case for the current study the HTC Vive HMD was used (https://www.vive.com/us/). The Vive (released in 2016) includes 2

external sensors for location detection, a head mounted display for displaying visual information, and over-ear speakers for emitting surround sound. The field of view provided by the Vive is approximately 110° (approximately 90° per eye). Each eye receives visual information from a separate 1080×1200 pixel display with a 90 Hz refresh rate. Participants respond to Stroop stimuli with an appropriate key press on a keyboard. The Stroop stimuli are presented one at a time. Once a participant responds to a Stroop stimulus the next is presented. In each of the Stroop conditions participants can experience up to 50 Stroop items, however each zone has a time limit before the next zone starts. If participants take a long time to respond to stimuli, they may not experience all 50 items in a zone before the next zone starts.

2.2.3 Analyses

Factor analyses were conducted to examine and compare the underlying constructs of the VRST and the ANAM. Analyses were performed using SPSS software version 27. Outcome variables used in the analyses consisted of mean reaction times and percentage of correct responses from each of the Stroop conditions (e.g., word reading). Factor analyses were conducted with both safe zone data from the VRST together and with the conditions separated due to differences between the safe zones and the ambush zones. Because the ANAM does not include a complex interference condition, data from the VRST complex interference condition was not included in the factor analysis. Participants with scores on variables greater than two standard deviations from the mean had those variable scores removed from the analyses (91 participants had data for all VRST conditions, 97 had data for safe zones, 102 had data for all ambush zones, and 98 had data for the ANAM Stroop). Normality of the data was assessed by examining skewness and kurtosis values and visually inspecting histograms. Normally distributed data can improve the performance of several types of extraction methods such as

principal axis factoring (PAF; Yong & Pearce, 2013).

Before performing the factor analyses, the appropriateness of factor analysis was evaluated by examining correlation matrixes (Tabachnick & Fidell, 2013), Kaiser-Meyer-Olkin Measure of Sampling Adequacy (KMO), and Bartlett's test of sphericity (Howard, 2016). KMO scores should be above 0.5, but 0.6 and above is more desirable (Howard, 2016). When performing the extraction during the analyses, PAF was performed (Ngure et al., 2015). Further, the correlation matrix was used as opposed to the variance-covariance matrix to increase interpretability of results (Yong & Pearce, 2013).

Several sources were used to determine the number of factors to extract including eigenvalues, scree plots, parallel analysis, and the minimum average partial (MAP) test. Direct oblimin (with delta set to 0) is an oblique rotation that was used in the current study for factor rotations (Osborne, 2015). Factor loadings in the current study were based on pattern matrices. Guidelines for assigning variables to factors based on factor loadings come from Howard (Howard, 2016). Participant factor scores from the Stroop tasks were examined for potential correlations between the underlying factors across Stroop tasks, using Pearson's *r*. Finally, differences between scores from the safe zones and ambush zones were examined via paired-samples tests. These tests assess differences within participants across measures.

2.3 Results

Descriptive statistics for percent of correct responses and mean response times to variables from the Stroop tasks can be seen in Table 2.2. Skewness, kurtosis, and histograms for the variables indicated that the variables were normally distributed.

Table 2.2

Descriptive Statistics

Variable	*M*	*SD*	Skewness	Kurtosis
VRST % correct WR Safe	93.65	4.47	-0.57	-0.30
VRST % correct CN Safe	93.22	5.26	-0.59	-0.28
VRST % correct SI Safe	92.40	5.79	-0.69	0.07
VRST % correct WR Ambush	92.04	4.98	-0.28	-0.57
VRST % correct CN Ambush	91.71	5.17	-0.57	-0.22
VRST % correct SI Ambush	91.28	6.17	-0.51	-0.46
VRST reaction time WR Safe	962.07	116.48	0.32	-0.43
VRST reaction time CN Safe	866.34	99.59	0.40	0.33
VRST reaction time SI Safe	1056.08	155.32	0.27	0.06
VRST reaction time WR Ambush	889.34	107.48	0.32	-0.12
VRST reaction time CN Ambush	869.39	92.57	0.27	-0.16
VRST reaction time SI Ambush	992.77	145.58	0.04	-0.42
ANAM % correct WR	97.04	3.64	-1.21	0.55
ANAM % correct CN	95.33	3.76	-0.38	-0.91
ANAM % correct SI	92.42	6.47	-1.12	1.02
ANAM reaction time WR	644.61	91.85	0.06	-0.44
ANAM reaction time CN	581.34	91.25	0.36	-0.61
ANAM reaction time SI	765.74	171.25	0.72	0.17

Note. ANAM = Automated Neuropsychological Assessment Metrics; VRST = Virtual Reality Stroop Task; M = mean; SD = standard deviation; WR = word reading; CN = color naming; SI = simple interference; % correct = percentage of correct responses to stimuli; reaction time is in milliseconds; Safe indicates scores from the safe zones of the VRST; Ambush indicates scores from the ambush zones of the VRST.

2.3.1 VRST Combined Results

The KMO (0.76) score which was above 0.6 indicated well defined factors (Howard, 2016) and Bartlett's test of sphericity, $\chi^2(66) = 715.31$, $p < .001$, indicated that underlying factors likely exist for correct responses and reaction times from the VRST. There were two factors with an eigenvalue greater than one for the VRST, which accounted for 58.31% of the total variance. Examination of the scree plot also indicated that two factors should be extracted (Figure 2.1).

The original MAP and revised MAP test indicated that two factors should be extracted (Table 2.3). Finally, parallel analysis indicated that possibly all factors could be extracted, however factors with initial eigenvalues less than 1 were not considered for extraction, Table 2.4.

Figure 2.1

Scree Plot VRST Combined Safe and Ambush Zone

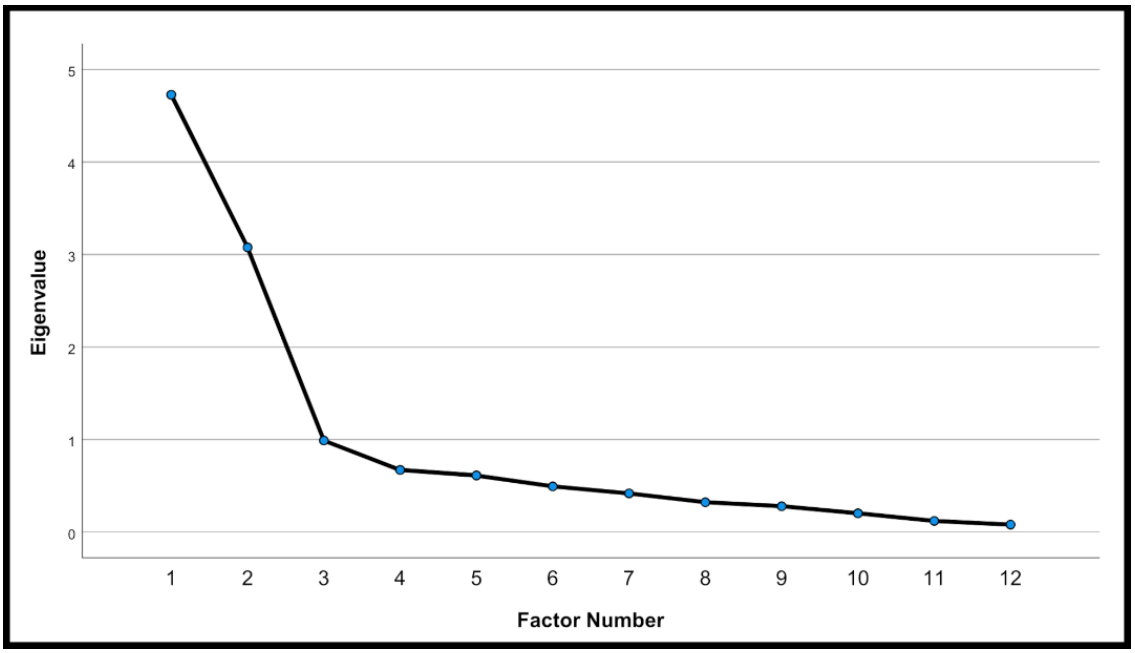


Table 2.3

MAP VRST Combined

Factor	Squared	4th power
0	0.169	0.061
1	0.139	0.032
2	**0.056**	**0.008**
3	0.058	0.010
4	0.075	0.018
5	0.105	0.030
6	0.135	0.052
7	0.165	0.065

(table continues)

Factor	Squared	4th power
8	0.222	0.109
9	0.358	0.210
10	0.462	0.358
11	1.000	1.000

Note. The minimum average partial is in bold.

Table 2.4

Randomly Generated Eigenvalues from Parallel Analysis VRST Combined

Root	PAF	
	Mean	95th Percentile
1	0.779	0.978
2	0.595	0.749
3	0.454	0.553
4	0.328	0.424
5	0.219	0.305
6	0.120	0.213
7	0.025	0.103
8	-0.062	-0.002
9	-0.135	-0.080
10	-0.211	-0.165
11	-0.290	-0.240
12	-0.363	-0.328

Note. PAF = principal axis factoring; means indicate the average eigenvalue of randomly generated data; 95th percentile indicates that fewer than 5 percent of randomly generated data has eigenvalues greater than the indicated value.

As can be seen in the pattern matrix (Factor loadings, Table 2.5) the first extracted factor was a measure of reaction time. The second extracted factor was a measure of percentage of correct responding. The correlation between these factors was weak, $r = .18$.

Table 2.5

PAF Loadings from VRST Combined

Variable	Factor 1	Factor 2
VRST % correct WR Safe	0.07	**0.65**
VRST % correct CN Safe	0.14	**0.73**
VRST % correct SI Safe	0.12	**0.62**
VRST % correct WR Ambush	-0.17	**0.67**
VRST % correct CN Ambush	-0.02	**0.79**
VRST % correct SI Ambush	-0.05	**0.74**
VRST reaction time WR Safe	**0.74**	0.00
VRST reaction time CN Safe	**0.75**	0.06
VRST reaction time SI Safe	**0.71**	-0.03
VRST reaction time WR Ambush	**0.92**	0.07
VRST reaction time CN Ambush	**0.86**	0.03
VRST reaction time SI Ambush	**0.85**	-0.08

Note. Principal axis factoring (PAF) factor loadings indicate values from pattern matrix; VRST = Virtual Reality Stroop Task; WR = word reading; CN = color naming; SI = simple interference; % correct = percentage of correct responses to stimuli; Safe indicates scores from the safe zones of the VRST; Ambush indicates scores from the ambush zones of the VRST.

2.3.2 VRST Safe Zones and Ambush Zones

KMO (0.62) scores were considered acceptable, KMO scores should be above 0.5, but 0.6 and above is more desirable, see Howard (2016) for review (Howard, 2016). Additionally, Bartlett's test of sphericity, $\chi^2(15) = 187.63$, $p < .001$, indicated that underlying factors likely exist for reaction times and percentage of correct responses from the VRST safe zones. Two factors had eigenvalues greater than one for the VRST safe zones, which accounted for 53.67% of the total variance. Examination of the scree plot also indicated that two factors should be extracted (Figure 2.2). The original MAP test suggested that no factors should be extracted but the revised MAP test indicated that two factors should be extracted, Table 2.6. Finally, parallel analysis indicated that possibly all factors should be extracted, however factors with eigenvalues

less than 1 were not considered for extraction, Table 2.7. After review of all the extraction rules two factors were selected for extraction.

Similar to the results when safe zones and ambush zones were combined, examination of the data from the safe zones revealed that the first extracted factor was a measure of reaction time, and the second extracted factor was a measure of percentage of correct responding, Table 2.8. The factors were only somewhat correlated with each other, $r = .33$.

The KMO score for the ambush zones (0.57) was above 0.5 and considered acceptable. Bartlett's test of sphericity, $\chi^2(15) = 315.63$, $p < .001$, indicated that underlying factors likely exist for percent correct responses and response time for VRST ambush zones. There were two factors with eigenvalues greater than one, which accounted for 63.31% of the total variance. Examination of the scree plot also indicated that two factors should be extracted (Figure 2.2). The original MAP and revised MAP test both indicated that one factor should be extracted, Table 2.9. Finally, parallel analysis indicated that all factors should be extracted, but again factors without eigenvalues greater than 1 were not considered for extraction, Table 2.10. Therefore, researchers extracted two factors.

As with the results from both the combined analysis and safe zones, the first extracted factor when examining ambush zone data was a measure of reaction time and the second extracted factor was a measure of percentage of correct responding, Table 2.11. The factors were uncorrelated with each other, $r = -.03$.

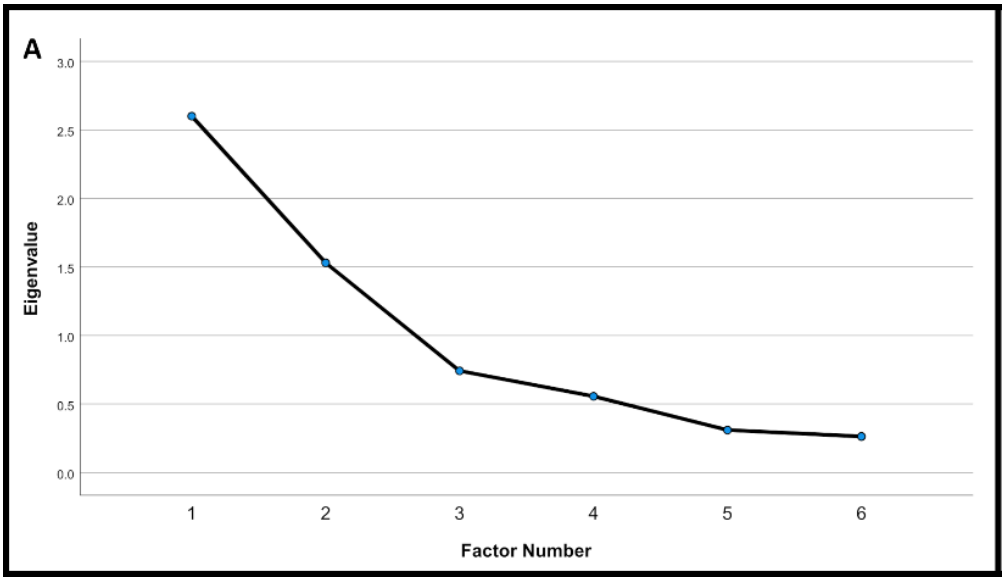
2.3.3 Throughput Assessment

To examine potential differences in performance between safe zones and ambush zones throughput scores were compared within participants. Results indicated that statistically significant differences existed between throughput scores within the word reading ($p < .001$),

color naming (p = .046), simple interference (p < .01), and complex interference conditions, p < .001. Throughput was found to increase within ambush zones compared to safe zones except for the color naming condition where, arousing stimuli tended to decrease performance, see Table 2.12.

Figure 2.2

Scree Plot VRST Safe Zones (A) and Ambush Zones (B)



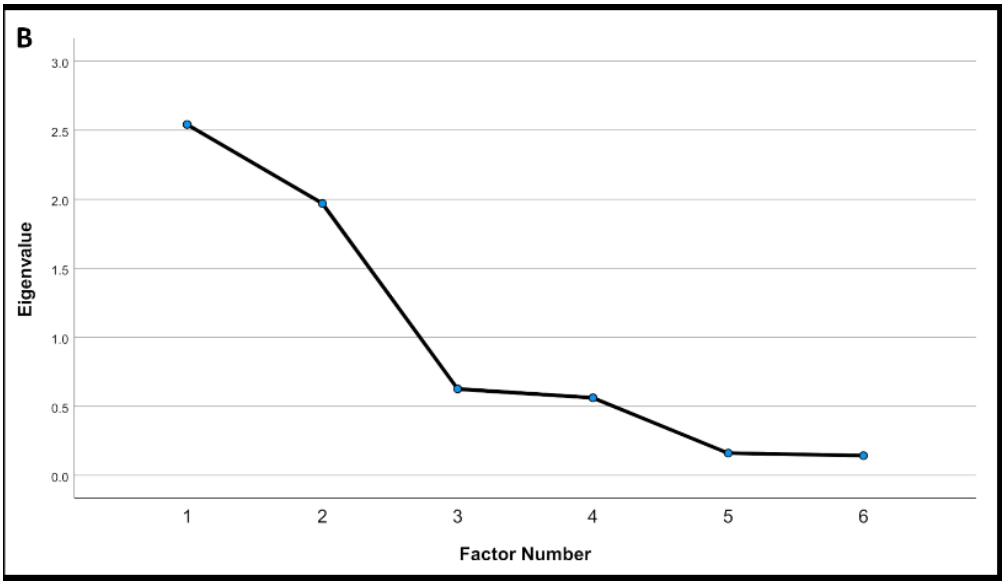


Table 2.6

MAP VRST Safe Zone

Factor	Squared	4th power
0	**0.138**	0.036
1	0.149	0.043
2	0.144	**0.033**
3	0.228	0.128
4	0.404	0.259
5	1.000	1.000

Note. The minimum average partial is in bold.

Table 2.7

Randomly Generated Eigenvalues from Parallel Analysis VRST Safe Zone

Root	PAF	
	Means	95th Percentile
1	0.402	0.559
2	0.220	0.335
3	0.082	0.171
4	-0.031	0.031
5	-0.124	-0.057
6	-0.250	-0.188

Note. PAF = principal axis factoring; means indicate the average eigenvalue of randomly generated data; 95th percentile indicates that fewer than 5 percent of randomly generated data has eigenvalues greater than the indicated value.

Table 2.8

PAF Loadings from VRST Safe Zones

Variable	Factor 1	Factor 2
VRST % correct WR Safe	-0.02	**0.70**
VRST % correct CN Safe	0.02	**0.72**
VRST % correct SI Safe	0.00	**0.68**

(table continues)

Variable	Factor 1	Factor 2
VRST reaction time WR Safe	**0.81**	-0.03
VRST reaction time CN Safe	**0.82**	0.03
VRST reaction time SI Safe	**0.64**	0.01

Note. Principal axis factoring (PAF) factor loadings indicate values from pattern matrix; VRST = Virtual Reality Stroop Task; WR = word reading; CN = color naming; SI = simple interference; % correct = percentage of correct responses to stimuli; reaction time is in milliseconds; Safe indicates scores from the safe zones of the VRST.

Table 2.9

MAP VRST Ambush Zone

Factor	Squared	4th power
0	0.170	0.075
1	**0.146**	**0.030**
2	0.163	0.039
3	0.241	0.137
4	0.410	0.264
5	1.000	1.000

Note. The minimum average partial is in bold.

Table 2.10

Randomly Generated Eigenvalues from Parallel Analysis VRST Ambush Zone

Root	PAF	
	Means	95th Percentile
1	0.422	0.622
2	0.242	0.404
3	0.098	0.181
4	-0.031	0.032
5	-0.142	-0.086
6	-0.254	-0.187

Note. PAF = principal axis factoring; means indicate the average eigenvalue of randomly generated data; 95th percentile indicates that fewer than 5 percent of randomly generated data has eigenvalues greater than the indicated value.

Table 2.11

PAF Loadings from VRST Ambush Zones

Variable	Factor 1	Factor 2
VRST % correct WR Ambush	-0.11	**0.63**
VRST % correct CN Ambush	0.13	**0.66**
VRST % correct SI Ambush	-0.02	**0.78**
VRST reaction time WR Ambush	**0.95**	0.11
VRST reaction time CN Ambush	**0.80**	0.04
VRST reaction time SI Ambush	**0.86**	-0.16

Note. Principal axis factoring (PAF) factor loadings indicate values from pattern matrix; VRST = Virtual Reality Stroop Task; WR = word reading; CN = color naming; SI = simple interference; % correct = percentage of correct responses to stimuli; reaction time is in milliseconds; Ambush indicates scores from the ambush zones of the VRST.

Table 2.12

Throughput Scores VRST

Throughput	*M*	*SD*
Safe zone WR	0.97	0.14
Ambush zone WR	1.03	0.14
Safe zone CN	1.07	0.12
Ambush zone CN	1.04	0.12
Safe zone SI	0.87	0.13
Ambush zone SI	0.91	0.16
Safe zone CI	0.75	0.14
Ambush zone CI	0.84	0.17

Note. WR = word reading; CN = color naming; SI = simple interference; CI = complex interference; Safe indicates scores from the safe zones of the VRST; Ambush indicates scores from the ambush zones of the VRST.

2.3.4 ANAM

KMO (0.70) scores indicated that factors were well defined (Howard, 2016) and Bartlett's test of sphericity, $\chi^2(15) = 227.69$, $p < .001$, indicated that underlying factors likely exist for percent correct responses and response time from the ANAM. There were two factors

with initial eigenvalues greater than one, which accounted for 55.81% of the total variance. Examination of the scree plot also indicated that one or two factors could be extracted (Figure 2.3). The original MAP and revised MAP test indicated that one factor should be extracted, Table 2.13. Finally, parallel analysis again indicated that all factors could be extracted. Given that the initial eigenvalue was greater than one, scree plot indicated more than one factor could be extracted, parallel analysis indicated that more than one factor could be extracted (Table 2.14) and the factors followed a similar structure to the VRST factors two factors were extracted. The first factor from the ANAM measured response time and the second factor measured percent correct responding. Factor loadings are shown in table 2.15. The factors were only somewhat correlated with each other, r = .32.

Figure 2.3

Scree Plot for ANAM Stroop

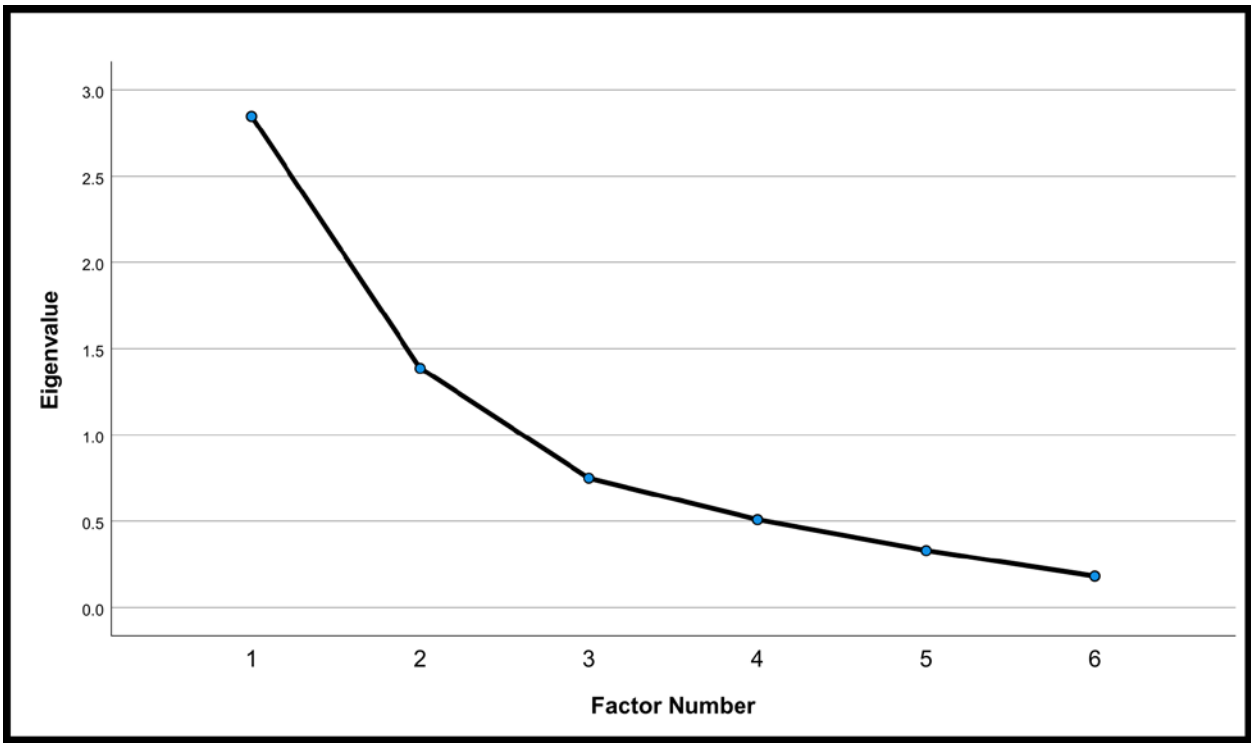


Table 2.13

MAP ANAM

Factor	Squared	4th power
0	0.166	0.059
1	**0.126**	**0.028**
2	0.139	0.039
3	0.222	0.152
4	0.539	0.413
5	1.000	1.000

Note. The minimum average partial is in bold.

Table 2.14

Randomly Generated Eigenvalues from Parallel Analysis ANAM

Root	PAF	
	Means	95th Percentile
1	0.395	0.582
2	0.226	0.337
3	0.082	0.193
4	-0.024	0.035
5	-0.136	-0.075
6	-0.246	-0.174

Note. PAF = principal axis factoring; means indicate the average eigenvalue of randomly generated data; 95th percentile indicates that fewer than 5 percent of randomly generated data has eigenvalues greater than the indicated value.

Table 2.15

PAF Loadings from ANAM

Variable	Factor 1	Factor 2
ANAM % correct WR	0.14	**0.60**
ANAM % correct CN	0.12	**0.65**
ANAM % correct SI	-0.12	**0.56**

(table continues)

Variable	Factor 1	Factor 2
ANAM reaction time WR	**0.78**	0.10
ANAM reaction time CN	**0.89**	0.10
ANAM reaction time SI	**0.82**	-0.11

Note. Principal axis factoring (PAF) factor loadings indicate values from pattern matrix; ANAM = Automated Neuropsychological Assessment Metrics; VRST = Virtual Reality Stroop Task; M = mean; SD = standard deviation; WR = word reading; CN = color naming; SI = simple interference; % correct = percentage of correct responses to stimuli; reaction time is in milliseconds.

2.3.5 Factor Correlations

Lastly, scores from the factors were created for each participant. These scores were then correlated to determine how related the factors are to each other, Table 2.16.

Table 2.16

Factor Score Correlations

Factor	1	2	3	4	5	6	7	8
1 VRST combined factor 1	1							
2 VRST combined factor 2	.20	1						
3 VRST safe zones factor 1	.89**	.25*	1					
4 VRST safe zones factor 2	.36**	.91**	.40**	1				
5 VRST ambush zones factor 1	.99**	.18	.84**	.34**	1			
6 VRST ambush zones factor 2	.07	.92**	.17	.70**	-.03	1		
7 ANAM factor 1	.45**	.23*	.48**	.30**	.47**	.14	1	
8 ANAM factor 2	.09	.50**	.16	.59**	.10	.39**	.43**	1

Note. * indicates $p < .05$; ** indicates $p < .01$; ANAM = Automated Neuropsychological Assessment Metrics; VRST = Virtual Reality Stroop Task.

CHAPTER 3

CLASSIFICATION OF PERFORMANCE IN THE VIRTUAL REALITY STROOP TASK USING MACHINE LEARNING

3.1 Purpose of Classifier Assessment

The current study was designed to examine ML strategies for use as performance classifiers. An important initial step is determining if the classification algorithms can accurately identify participant performance (McMahan et al., 2021). Examination and classification of participant performance can then be used to allow the VRST to adapt to users. Algorithms used in the current study (NB, SVM, and kNN) have been previously examined in other contexts, finding that no one classifier outperformed others in all classification tasks (McMahan et al., 2021; Parsons et al., 2022). Multiple ML approaches were used in the current paper and could be considered supervised ML (Ippolito , 2021; Mosavi et al., 2018). Supervised ML techniques require the researcher to select which variables will be used for classification (e.g., Researchers specifies that participants are labeled as higher or low performers based on throughput and outcome variables from VRST are used to classify group membership; Ippolito, 2021). Unsupervised learning on the other hand, creates groups/categories based on input data. For example, k-means clustering attempts to find different clusters or groups already existing within the data (Ippolito, 2021). Further, each of these approaches have several hyper-parameters which are selected by the researcher (Luo, 2016). Hyper-parameters are values or aspects of the algorithm that can be manipulated to influence how the ML algorithm performs (Luo, 2016). For NB the hyper-parameter manipulated in the current study was the use of a gaussian or kernel NB. For SVM various types of kernels were also examined. Finally, for kNN hyper-parameters included the number of neighbors used as well as distance formula. This study examined the

performance of these ML algorithms using several performance indicators such as percentage of correct classifications, precision, sensitivity, and area under the curve. Participants were classified as high and low performers within safe zones, ambush zones, and combined performance from the VRST.

3.2 Methods

3.2.1 Participants

The current study used a sample of college aged participants in the southwestern United States (*N* = 157; see Table 3.1 for demographics).

Table 3.1

*Demographics (*N *= 157)*

Characteristic	Mean (SD)	%
Age	23.17 (6.79)	
Men		54.78
Women		45.22
Black		14.01
Asian		8.92
Hispanic		16.56
Native American		1.91
Non-Hispanic white		55.41
Other/Multiracial		3.18

The sample consisted of slightly more men than women (54.78% men), and participants primarily identified as non-Hispanic white (55.41%). For outlier removal, scores on variables greater than 2 standard deviations from the mean had those variable scores removed from the analysis. Removal of outliers can improve classification performance of multiple types of classification algorithms (Hautamäki et al., 2005). NB and kNN classifiers can utilize incomplete cases, but SVMs require participants to not be missing data. Therefore, SVMs were conducted

on 86 participants within the combined analysis, 103 participants in the safe zones, and 106 participants in the ambush zones

3.2.2 Materials: VRST (Virtual Reality Stroop Task)

As described above, this study used the high mobility multipurpose wheeled vehicle (HMMWV) version of the VRST, Figure 1.2. The VRST-HMMWV is a virtual environment that simulates a middle eastern countryside. Participants are situated in the driver seat of the HMMWV while it is traversing down a desert road. Participants are not required to drive the vehicle and the simulated HMMWV travels along a pre-determined path. Within the VRST participants experience two zone types: safe and ambush. During the ambush zones the participant encounters arousing stimuli which includes explosions, shouting, and sounds of gunfire. Safe zones do not include the arousing stimuli and have few stimuli other than the HMMWV driving along the road. The zone participants experienced first, safe or ambush, was counterbalanced.

The VRST has 4 Stroop conditions that are each experienced in both safe and ambush zones, for a total of 8 conditions. The Stroop conditions presented stimuli to the participant in the center of the vehicle's windshield starting with color naming, which required the user to respond to colored Xs (XXXX). The second condition is word reading, participants selected the appropriate key on the keyboard that represents the color words (RED, BLUE, GREEN). The third condition was a simple interference condition where participants indicated the font color of the color words (BLUE). The fourth condition, the complex interference, added an additional degree of difficulty. In the complex interference condition, the Stroop stimuli appeared in various locations of the windshield instead of the stimuli appearing in the middle of the windshield, as was the case for the other conditions, see Figure 1.2.

The participants were immersed into the VRST using the HTC Vive head mounted display (HMD) (https://www.vive.com/us/). While using the HTC Vive, participants can turn their heads to look around the environment and audio was presented via headphones. Additionally, the VRST provided participants with tactile feedback via a bass shaker speaker underneath the participant's chair. Participants responded to Stroop stimuli by pressing the appropriate keys on the computer keyboard. Once a participant responds to a Stroop stimulus the next stimulus is presented. Participants can experience up to 50 items in each condition within each zone, but there is a limited amount of time to respond to stimuli in each condition therefore participants may not experience all 50 items.

3.2.3 Procedures

The current analysis compares the performance of multiple classification algorithms and hyper-parameters including SVM, NB, and kNN using MATLAB (version 2021a; The MathWorks, Inc.). Multiple measures of performance were used, and all measures had skewness and kurtosis values within +/- 3, indicating that the variables could be considered normally distributed. The VRST collects speed (reaction times for correct responses) and accuracy scores from all 4 Stroop conditions in both safe and ambush zones. As indicated by Thorne (2006), participants may use different strategies when performing timed tests, some participants may prioritize accurate responding while others may favor faster responding. Throughput, which is a measure of accuracy divided by time was used because it accounts for tradeoffs participants make between speed and accuracy (Thorne, 2006). Therefore, throughput was used as a measure of overall performance, which takes into account both accuracy and speed. A median split was conducted to classify participant performance, participants were considered high performers if their scores were above the median or low performers if their scores were below the median. In

all the classifiers, a 10-fold cross validation was used to improve the algorithms' ability to accurately classify new data. According to Jung (2018) a k-fold cross-validation is a leave-one-out cross-validation method. The dataset is first divided into k approximately equal sized datasets, the current study used a k of 10. The k-1 folds (i.e., 9 subsets of data) are used to train the classifier and the fold left out is used for testing classifier performance. This process iterates k times so that all k folds are used as validation data once.

3.2.3.1 Naïve Bayes

NB is based on Bayes theorem, which states that the posterior probability (i.e., probability of participant's membership in a particular group) is based on conditional probability (i.e., probability of membership in a particular group based on current data) weighted by previous knowledge (i.e., prior probability), divided by the probability of observing membership in a group. NB algorithms are called naïve because they treat each variable used in classification as independent from each other. NB algorithms tend to be less computationally intensive than other classifiers in terms of implementation and training (Vural, & Gök, 2017). During classification, likelihood scores are produced and compared for each possible category. Membership to that category is based upon which group the item to be classified is most likely to belong (i.e., has the greatest likelihood score). Our analysis used two versions of NB algorithms, the first was a typical gaussian NB where the predictors are assumed to be normally distributed. The second was kernel NB, where fewer assumptions are required, the classifier creates separate estimates for each predicted class based on training data rather than assuming a normal distribution.

3.2.3.2 Support Vector Machine

SVMs may be used for classification and can be understood via four concepts: the

separating hyperplane, the maximum-margin hyperplane, the soft margin, and the kernel function (Noble, 2006). The separating hyperplane is a higher-dimensional plane which separates the data into distinct groups. The maximum-margin hyperplane refers to the plane created when the SVM maximizes the distance between itself and the nearest expression vector, the distance is known as the margin (Noble, 2006). It is assumed that the greater the margin the better the SVM will perform (Bhavsar & Panchal, 2012). Often, real data does not neatly fit into distinct groups, therefore the soft margin is introduced. As reviewed by Noble (2006), the soft margin is a user-specified parameter which determines how much misclassification is acceptable. The final aspect of the SVM is the kernel function, which is a mathematical technique that allows data with fewer dimensions to be treated as higher dimensional data. Several variations of SVM kernel algorithms exist and are used to maximize the margin including linear, polynomial, and radial basis to influence the shape of the separating hyperplane (McMahan et al., 2021). The kernel uses dot products to examine relationships between points within the dataset as if it were transformed to a higher dimension. Once data is examined in a higher dimensional space, a hyperplane may be able to separate the data into categories more accurately than points, lines, or lower-dimensional planes. This study examines performance of linear, polynomial, and various Gaussian kernel SVM algorithms.

3.2.3.3 k-Nearest Neighbor

kNN algorithms use a system similar to voting to classify data. First, the k in kNN indicates the number of other closest datapoints which are used in the classification. A k of 3 indicates the three closest datapoints are used, while k = 10 indicates the 10 closest datapoints are used for classifications. The closeness of datapoints is based on distance calculations for the included predictor variables. Non-weighted kNN algorithms simply use majority rule, for

instance if k = 10 and the 6 closest participants based on predictor scores were classified as high performers, the current participant would be classified as a high performer. However, kNN algorithms have become more sophisticated by including weights that are based on distances. In weighted kNN algorithms the more similar the surrounding cases are to the current data point (i.e., closer in distance) the more influence they have on the classification. There are multiple algorithms for weight based on distance, this study implements Euclidean distances for the weights for fine, medium, coarse, and weighted kNN, cosine for cosine kNN, and Minkowski for cubic kNN.

3.2.3.4 Classifier Assessment

When assessing machine algorithm performance, multiple metrics should be used such as area under the curve (AUC), correct classification rate, and specificity (Beunza et al., 2019; see Table 3.2 for a list of classification metrics used). Many of these metrics originated from signal detection theory and have been used for evaluation of classifiers (Flach, 2016). The metrics are often reported from the optimal classification threshold during training, as well as assessing generalizability, classifier comparisons, and for selecting optimal solutions; see Hossin & Sulaiman (2015) for review. When examining performance of binary classification (i.e., two groups) performance can be evaluated using number of correct and incorrect classifications as well as various other metrics based on these correct and incorrect classifications. Briefly, classes are labeled as zero or one (in the present study high performers were labeled as one and low performers were labeled as zero). True positives (TP) are the number of classifications where participant was in class one and classifier identified participant as class one. False negatives (FN) the number of classifications where participant was in class one and classifier identified participant as class zero. False positives (FP) the number of classifications where participant was

in class zero and classifier identified participant as class one. True negatives (TN) the number of classifications where participant was in class zero and classifier identified participant as class zero. Multiple performance metrics are based on formulas using these values.

Table 3.2

Classification Metrics

Metric	Formula	Description
TP		Correct positive classifications
FN		Incorrect negative classifications
FP		Incorrect positive classifications
TN		Correct negative classifications
Correct rate (accuracy)	$\frac{\mathrm{TP}+\mathrm{TN}}{\mathrm{TP}+\mathrm{FP}+\mathrm{TN}+\mathrm{FN}}$	Ratio of correct classifications to total classifications
Sensitivity	$\frac{TP}{TP+FN}$	Measure of positive classification performance
Specificity	$\frac{TN}{TN+FP}$	Measure of negative classification performance
Precision	$\frac{TP}{TP+FP}$	Probability positive classifications were correct
AUC	$\frac{S_p - n_p(n_n+1)/2}{n_p n_n}$	Measure of overall performance of a classifier at multiple thresholds

Note: TP = True positives; FN = False negatives; FP = False positives; TN = True negatives; AUC = Area under the curve; S_p indicates the sum of all positive examples ranked; n_p indicates the number of positive examples; n_n indicates the number of negative examples.

AUC is an indication of how well the classifier performs at different cutoff thresholds. Mandrekar (2010) gives approximate interpretations for AUC values, stating that an AUC of 0.5 indicates that the classifier has no diagnostic value and AUC of 0.7 to 0.8 is acceptable. Additionally, Mandrekar (Mandrekar, 2010) states that AUCs of 0.8 to 0.9 are excellent and classifiers with AUCs greater than 0.9 are outstanding. Often the AUC is visualized with a receiver operator characteristic (ROC) curve. ROC curves often display the true positive rate

(i.e., sensitivity) on the y axis and the false positive rate (i.e., 1-specificity) along the x axis, for various cutoff thresholds (Jiao & Du, 2016). These cutoff thresholds can be used to minimize false positive rates or to maximize the number of classifications for a desired outcome. For instance, it may be more beneficial in disease screening to decrease total classification accuracy by increasing false positive rates if it will lead to more true positive classifications and getting care to people who may need it.

3.3 Results

3.3.1 Overall Performance

Throughput from all Stroop conditions within both safe and ambush zones were used to identify participants as high or low performers. Predictors included percentage of correct responses and reaction times from all Stroop conditions. Descriptive statistics for the predictors are shown in Table 3.3. Participants completed a battery of tests in addition to the VRST.

Table 3.3

Descriptive Statistics for Predictor Variables

Predictor name	Mean (SD)	Min	Max
Safe WR % correct	93.9(4.15)	85.00	100.00
Safe WR reaction time	935.13(138.07)	617.72	1275.44
Safe CN % correct	93.5(4.56)	84.00	100.00
Safe CN reaction time	857.15(96.53)	644	1105.51
Safe SI % correct	91.75(6.36)	73.47	100.00
Safe SI reaction time	1056.91(169.17)	681.82	1453.91
Safe CI % correct	89.32(6.87)	72.09	100.00
Safe CI reaction time	1170.01(208.02)	745.78	1804.83
Ambush WR % correct	91.97(5.04)	80.00	100.00
Ambush WR reaction time	881.88(111.54)	671.68	1174
Ambush CN % correct	91.96(5.07)	80.00	100.00

(table continues)

Predictor name	Mean (SD)	Min	Max
Ambush CN reaction time	863.9(103.47)	662.1	1122.28
Ambush SI % correct	90.03(6.5)	71.43	100.00
Ambush SI reaction time	1014.25(149.61)	690.98	1329.56
Ambush CI % correct	88.16(7.99)	69.77	100.00
Ambush CI reaction time	1034.67(160.01)	680.84	1392.64

Note. WR = word reading; CN = color naming; SI = simple interference; % correct = percentage of correct responses to stimuli; reaction times are in milliseconds; SD = standard deviation; Min = minimum; Max = maximum.

When using all the data from the VRST cubic SVMs had the greatest percentage of correct classifications, however NB algorithms and SVMs other than fine Gaussian SVM also performed at similar levels; Figures 3.1. Additionally, many of the algorithms had AUC values which were considered good; see Table 3.4. Similar to ambush zone classification performance kNN algorithms tended to struggle when classifying participants as high performers as can be seen with poor sensitivity (Figure 3.2) and poor precision, Figure 3.3. Coarse kNN and fine Gaussian SVM labeled all participants as high performers, which led to poor scores for both sensitivity and precision.

Table 3.4

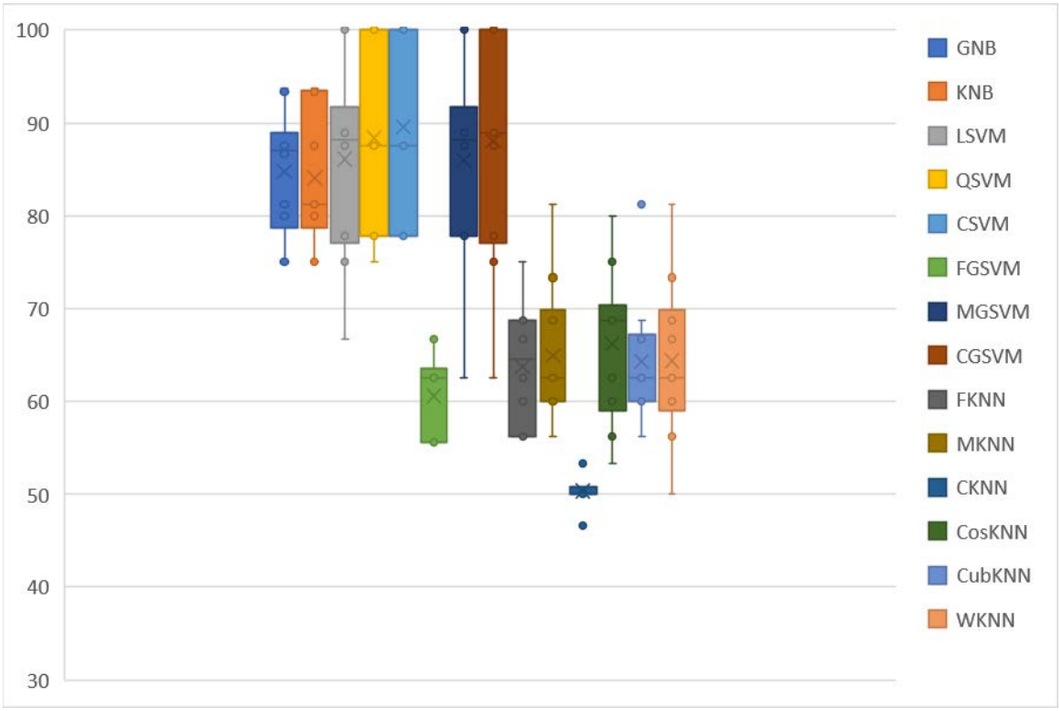
Classification Metrics for Combined Performance

	GNB	KNB	LSVM	QSVM	CSVM	FGSVM	MGSVM	CGSVM	FKNN	MKNN	CKNN	CosKNN	CubKNN	WKNN
N	157	157	86	86	86	86	86	86	157	157	157	157	157	157
TP	65	63	29	29	28	0	30	32	24	24	0	30	24	23
FN	11	10	7	5	3	0	8	8	3	1	0	5	2	1
FP	13	15	5	5	6	34	4	2	54	54	78	48	54	55
TN	68	69	45	47	49	52	44	44	76	78	79	74	77	78
Correct Rate (%)	84.71	84.08	86.05	88.37	89.53	60.47	86.05	88.37	63.69	64.97	50.32	66.24	64.33	64.33
Sensitivity	0.83	0.81	0.85	0.85	0.82	0.00	0.88	0.94	0.31	0.31	0.00	0.38	0.31	0.29
Specificity	0.86	0.87	0.87	0.90	0.94	1.00	0.85	0.85	0.96	0.99	1.00	0.94	0.97	0.99
Precision	0.83	0.81	0.85	0.85	0.82	0.00	0.88	0.94	0.31	0.31	0.00	0.38	0.31	0.29
AUC	0.90	0.90	0.97	0.96	0.96	0.51	0.95	0.97	0.82	0.95	0.36	0.96	0.95	0.97

Note: N = sample size; TP = True positives; FN = False negatives; FP = False positives; TN = True negatives; AUC = Area under the curve; GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.1

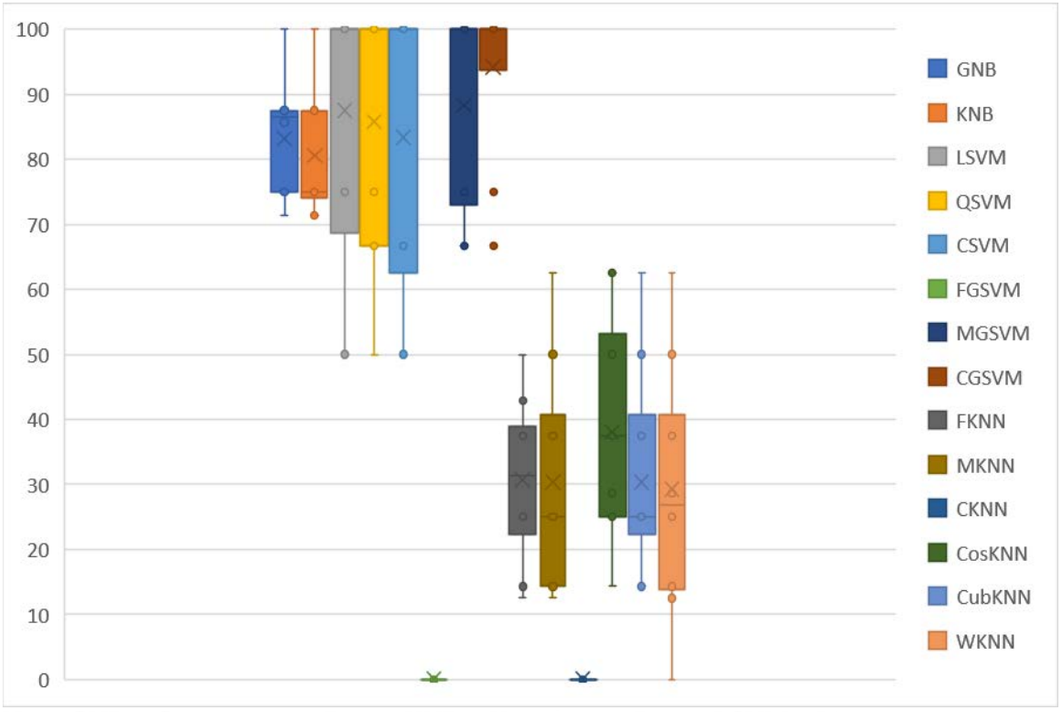
Accuracy for Classifications Combined Performance



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.2

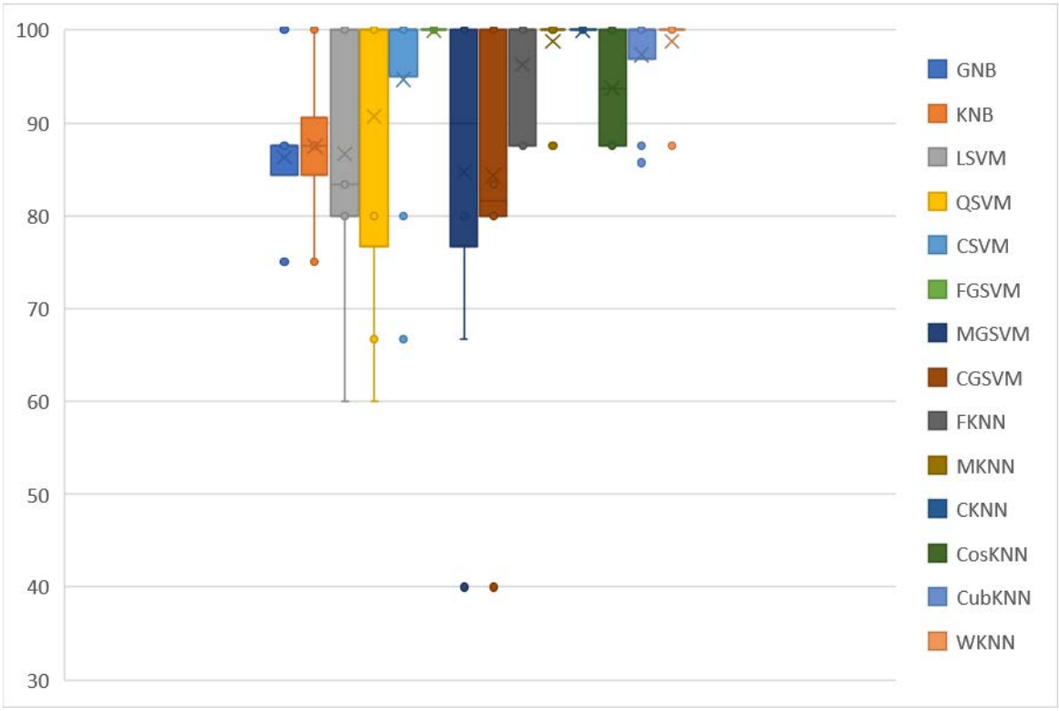
Sensitivity for Classifications Combined Performance



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.3

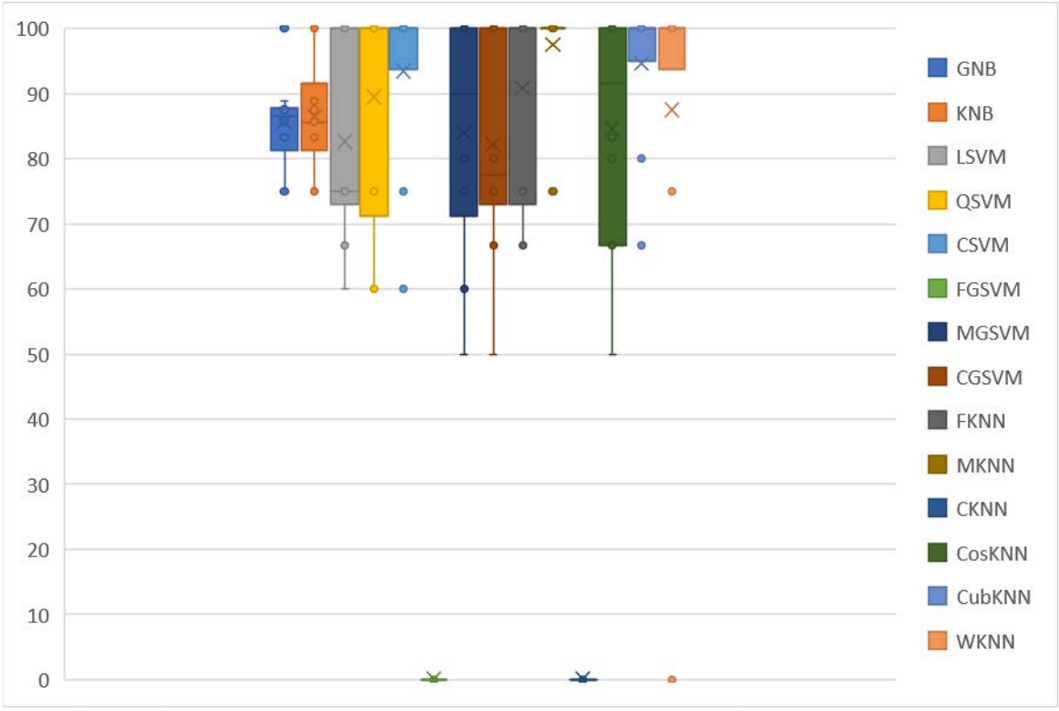
Specificity for Classifications Combined Performance



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.4

Precision for Classifications Combined Performance



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

3.3.2 Safe Zones

For safe zone performance throughput from all Stroop conditions within safe zones was calculated to identify participants as high or low performers based on percentage of correct responses and reaction times from all safe zone Stroop conditions, see Table 3.3 for descriptive statistics for predictors. There were 157 participants used to train the classifiers, however the number of cases used for SVM algorithms was reduced to 103 because the classifier requires complete cases to perform classifications, number of cases used can be seen in Table 3.5.

Results indicate that Gaussian NB and kernel NB had the greatest percentage of correct classifications (i.e., correctly identifying user as high or low performer). Gaussian NB and kernel NB had accuracies of 81.53% and 82.17% respectively. The linear, medium Gaussian, and course Gaussian SVMs had acceptable accuracies (i.e., >70% correct classifications; Figure 3.5). Many of the classifiers had AUCs of .91 indicating excellent classification performance at a range of cutoff thresholds. However, fine Gaussian SVM, and fine and coarse kNN performed poorly with correct classification rates close to chance approximately 50% correct classification rates. While many of the classifiers were able to classify low and high performers well, both fine Gaussian SVM and coarse kNN performed poorly when classifying positive cases with low sensitivity (Figure 3.6) precision scores (Figure 3.8), coarse kNN simply labeled all participants as high performers; see Table 3.5 and Figures 3.6 and 3.8. Box plots (Figures 3.5 - 3.8) show additional classification characteristics of the ML algorithms.

Table 3.5

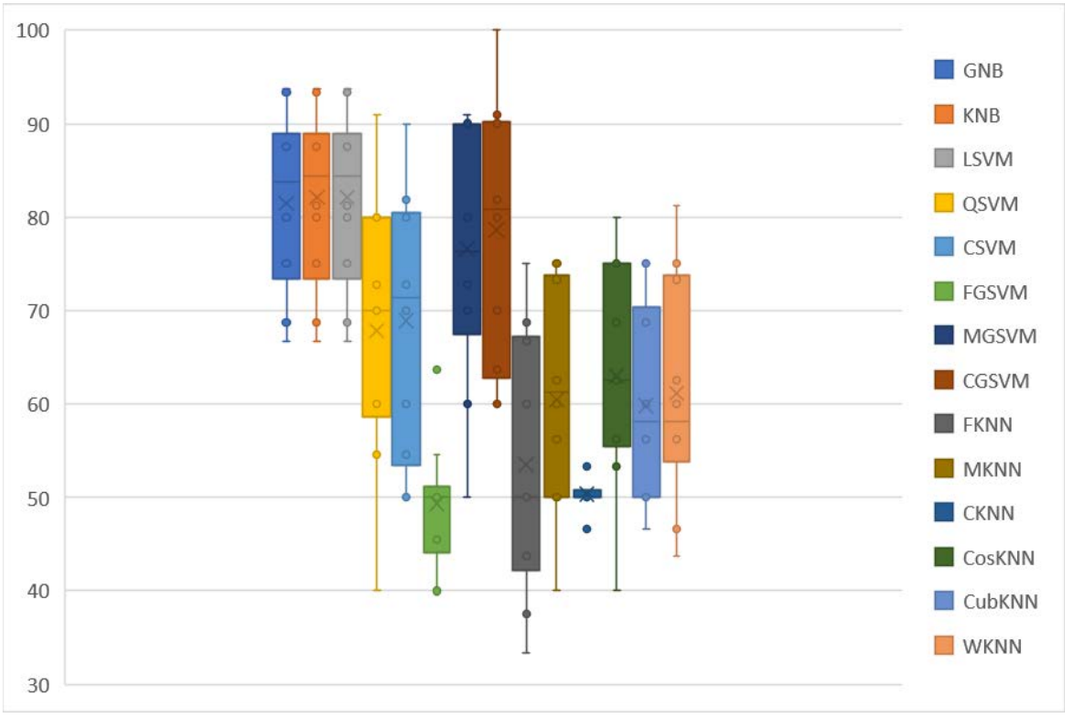
Classification Metrics for Safe Zone Performance

	GNB	KNB	LSVM	QSVM	CSVM	FGSVM	MGSVM	CGSVM	FKNN	MKNN	CKNN	CosKNN	CubKNN	WKNN
N	157	157	103	103	103	103	103	103	157	157	157	157	157	157
TP	64	62	37	33	30	2	40	40	22	32	0	40	31	31
FN	15	12	16	17	13	5	15	13	17	16	0	20	16	14
FP	14	16	12	16	19	47	9	9	56	46	78	38	47	47
TN	64	67	38	37	41	49	39	41	62	63	79	59	63	65
Correct Rate (%)	81.53	82.17	72.82	67.96	68.93	49.51	76.70	78.64	53.50	60.51	50.32	63.06	59.87	61.15
Sensitivity	0.82	0.79	0.76	0.67	0.61	0.04	0.82	0.82	0.28	0.41	0.00	0.51	0.40	0.40
Specificity	0.81	0.85	0.70	0.69	0.76	0.91	0.72	0.76	0.78	0.80	1.00	0.75	0.80	0.82
Precision	0.82	0.79	0.76	0.67	0.61	0.04	0.82	0.82	0.28	0.41	0.00	0.51	0.40	0.40
AUC	0.87	0.86	0.82	0.77	0.80	0.62	0.84	0.85	0.57	0.79	0.43	0.78	0.76	0.79

Note: N = sample size; TP = True positives; FN = False negatives; FP = False positives; TN = True negatives; AUC = Area under the curve; GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.5

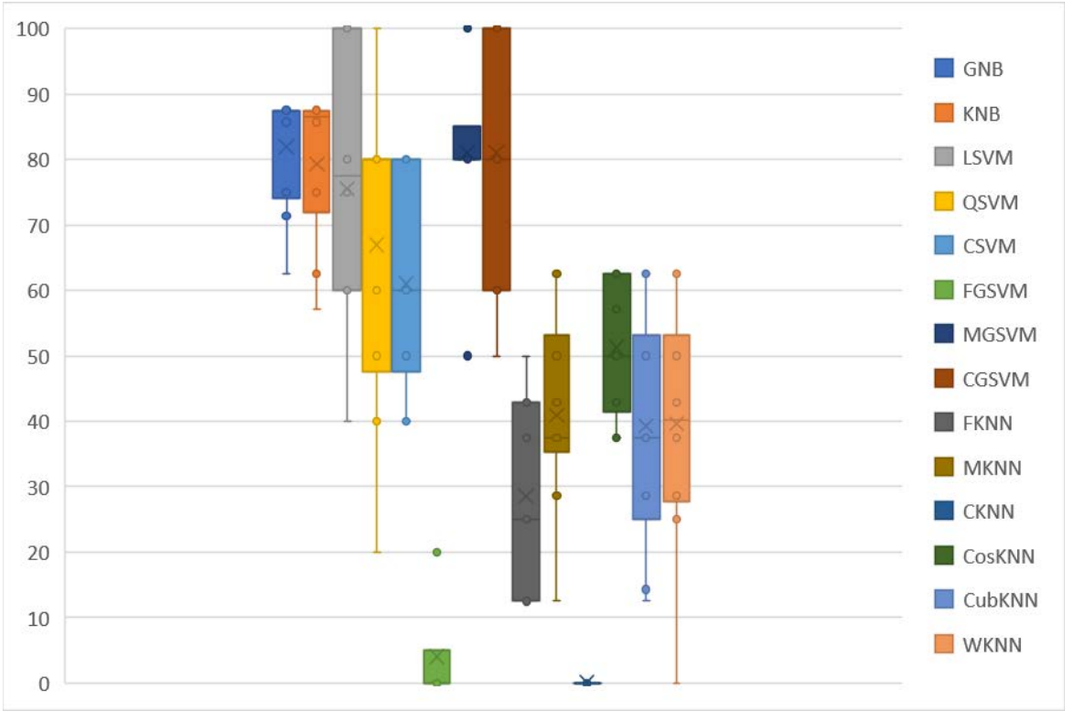
Accuracy for Classifications in Safe Zones



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.6

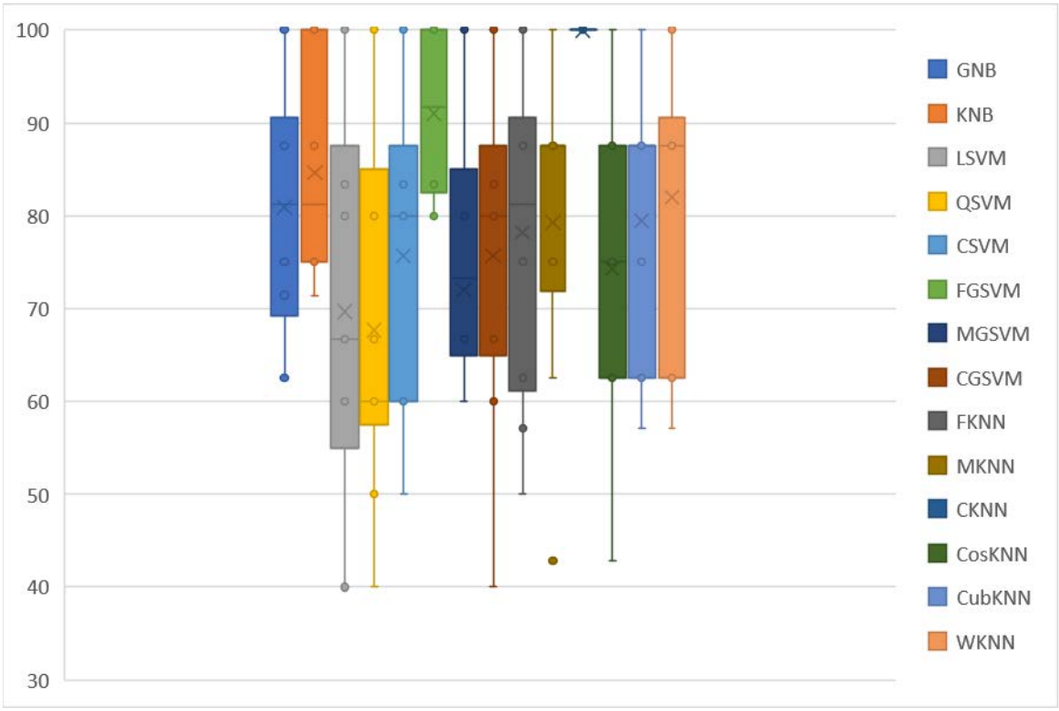
Sensitivity for Classifications in Safe Zones



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.7

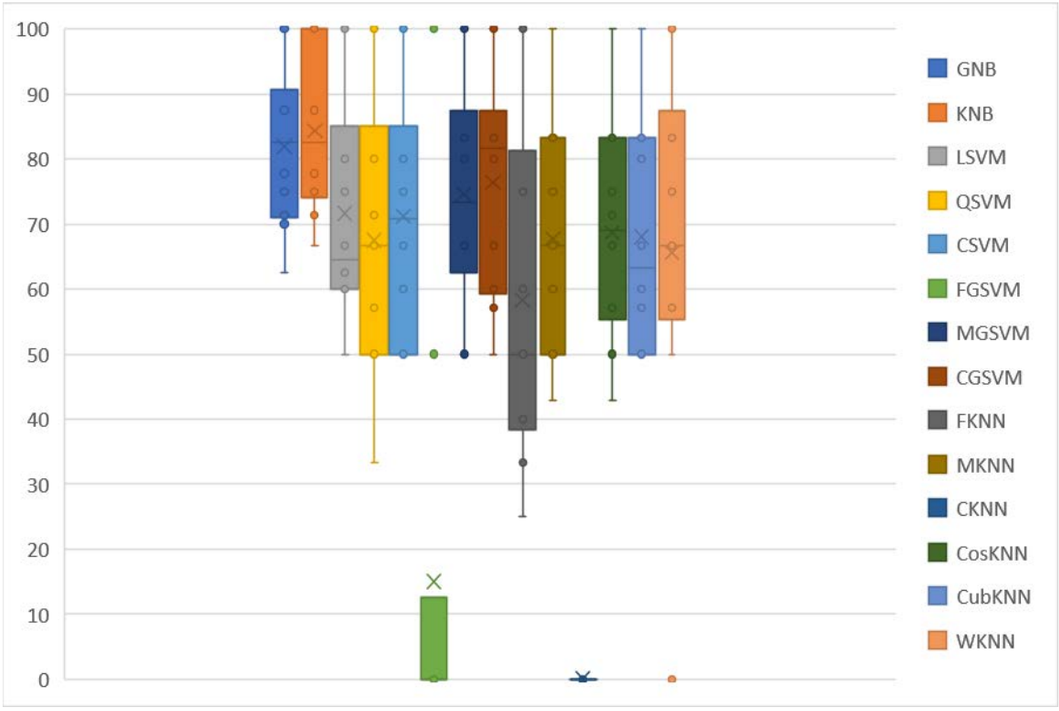
Specificity for Classifications in Safe Zones



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.8

Precision for Classifications in Safe Zones



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

3.3.3 Ambush Zones

Similar to the safe zone performance, throughput from all Stroop conditions within ambush zones was used to identify participants as high or low performers, with predictors including percentage of correct responses and reaction times from all ambush zone Stroop conditions, for descriptive statistics for predictors see Table 3.3. There were 156 participants used to train the classifiers (one participant's data was completely removed for ambush zones due to short reaction times and low accuracy). Additionally, SVMs were conducted on data from 106 participants who did not have any missing data.

Results indicate that medium Gaussian SVM had the greatest overall percentage of correct classifications (Figure 3.9) and greatest AUC. When predicting ambush zone performance classifiers performed better compared performance in safe zones. Only one algorithm had a correct classification rate which would be considered unacceptable (i.e., <70% correct classifications; see table 3.6). In general, kNN algorithms seemed to struggle with classifying users as high performers, often classifying too many participants as high performers when they should not, see Table 3.6 for false positive rate and Figures 3.10 and 3.12. As was the case for safe zones, the coarse kNN simply labeled all participants as low performers, which is why the classifier performed only slightly better than chance. Further, sensitivity and precision scores were quite low for both coarse kNN and fine Gaussian SVM (Figure 3.12), indicating these classifiers performed poorly when classifying positive cases.

Table 3.6

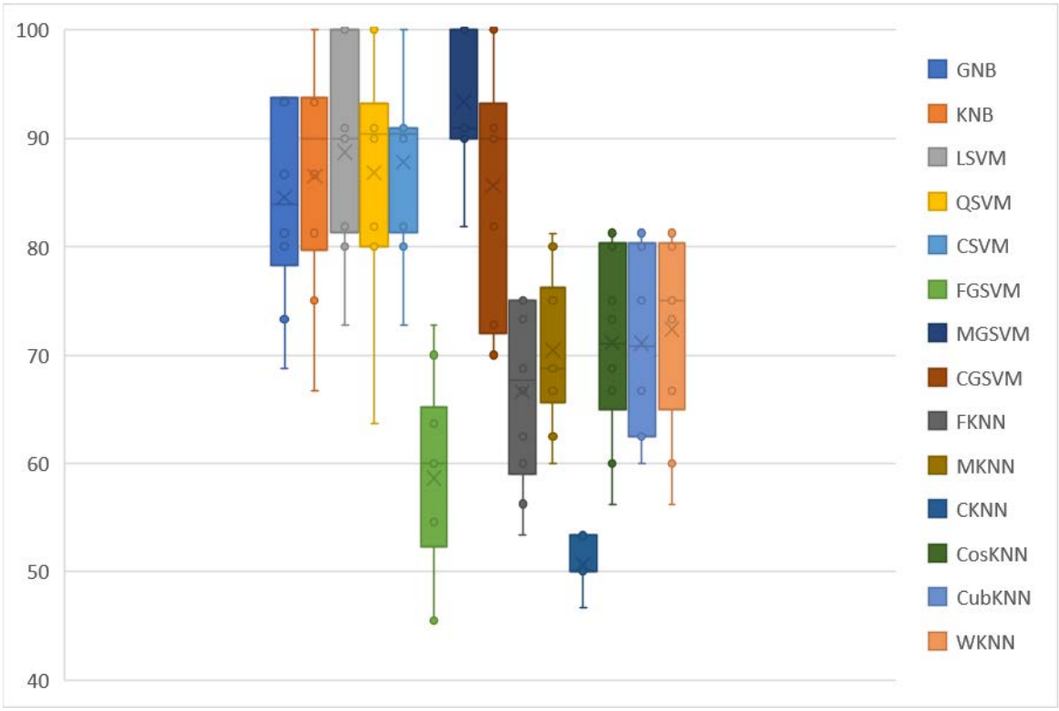
Classification Metrics for Ambush Zone Performance

	GNB	KNB	LSVM	QSVM	CSVM	FGSVM	MGSVM	CGSVM	FKNN	MKNN	CKNN	CosKNN	CubKNN	WKNN
N	156	156	106	106	106	106	106	106	156	156	156	156	156	156
TP	63	62	39	35	34	2	39	39	31	33	0	38	34	36
FN	10	6	7	5	3	2	2	10	6	2	0	6	2	2
FP	14	15	5	9	10	42	5	5	46	44	77	39	43	41
TN	69	73	55	57	59	60	60	52	73	77	79	73	77	77
Correct Rate (%)	84.62	86.54	88.68	86.79	87.74	58.49	93.40	85.85	66.67	70.51	50.64	71.15	71.15	72.44
Sensitivity	0.82	0.81	0.89	0.80	0.77	0.05	0.89	0.89	0.40	0.43	0.00	0.49	0.44	0.47
Specificity	0.87	0.92	0.89	0.92	0.95	0.97	0.97	0.84	0.92	0.97	1.00	0.92	0.97	0.97
Precision	0.82	0.81	0.89	0.80	0.77	0.05	0.89	0.89	0.40	0.43	0.00	0.49	0.44	0.47
AUC	0.90	0.90	0.93	0.92	0.88	0.83	0.94	0.92	0.80	0.92	0.42	0.93	0.93	0.93

Note: N = sample size; TP = True positives; FN = False negatives; FP = False positives; TN = True negatives; AUC = Area under the curve; GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.9

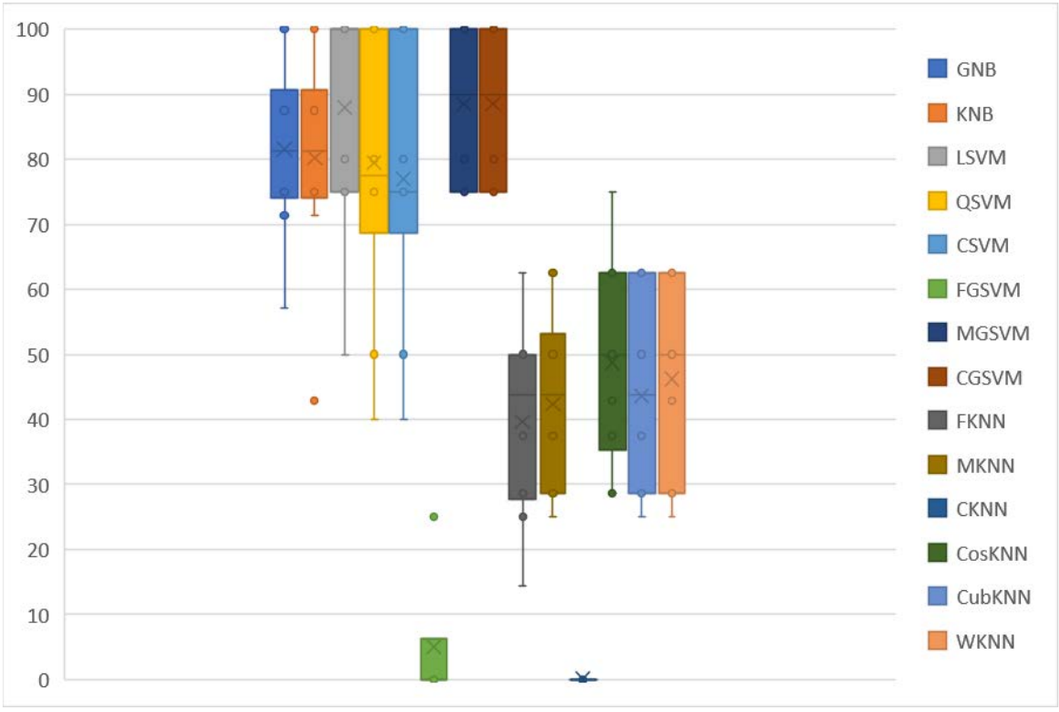
Accuracy for Classifications in Ambush Zones



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.10

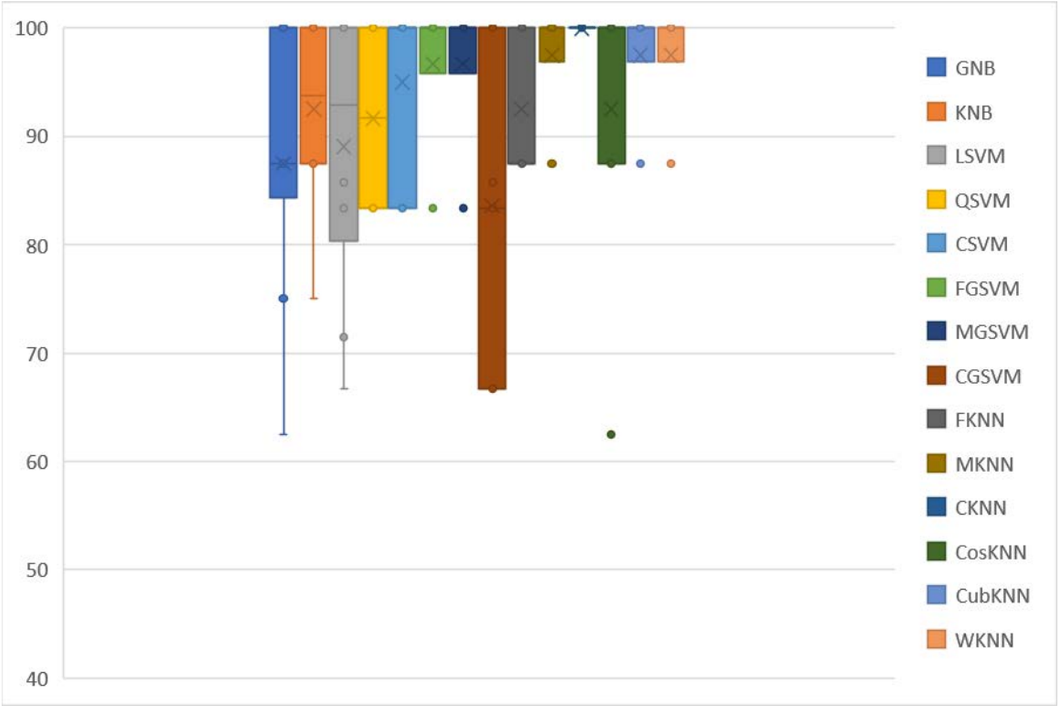
Sensitivity for Classifications in Ambush Zones



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.11

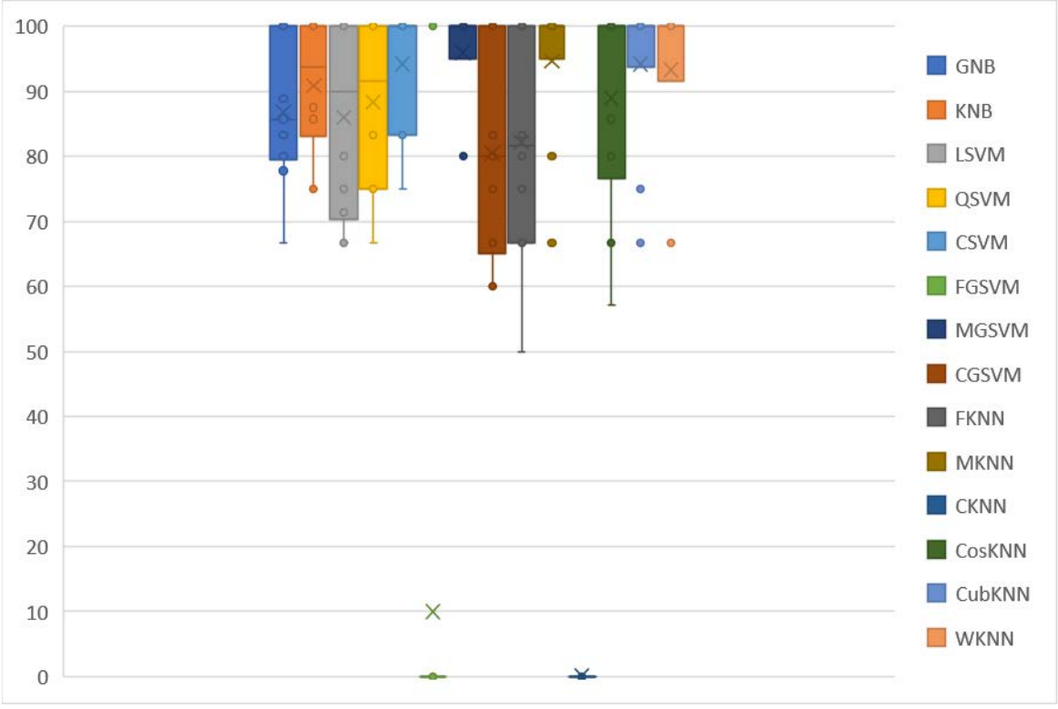
Specificity for Classifications in Ambush Zones



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

Figure 3.12

Precision for Classifications in Ambush Zones



Note: GNB = Gaussian naïve Bayes; KNB = Kernel naïve Bayes; LSVM = Linear support vector machines; QSVM = Quadratic support vector machines; CSVM = Cubic support vector machines; FGSVM = Fine Gaussian support vector machines; MGSVM = Medium Gaussian support vector machines; CGSVM = Coarse Gaussian support vector machines; FkNN = Fine k-Nearest Neighbor; MkNN = Medium k-Nearest Neighbor; CkNN = Coarse k-Nearest Neighbor; CoskNN = Cosine k-Nearest Neighbor; CubkNN = Cubic k-Nearest Neighbor; WkNN = Weighted k-Nearest Neighbor.

CHAPTER 4

DISCUSSION

4.1 Overview

The current study set out to accomplish two things: 1) provide additional validation for a higher dimensional Stroop task, the VRST, by comparing its factor structure the factor structure of a lower dimension Stroop task, the ANAM Stroop task. 2) investigate several machine learning algorithms and their hyper-parameters for the classification of participant performance for the creation of an adaptive version of the VRST.

The study examined the factor structure of the VRST and compared the results with a low dimensional version of the Stroop task the ANAM Stroop task. The study also sought to examine potential differences in factor structures between safe zones and ambush zones within the VRST. We found that when using the percentage of correct responses and reaction times from both the VRST and ANAM, two-factor solutions were obtained. These factors tended to be related to either accuracy of responses to Stroop stimuli or time taken to respond to the Stroop stimuli.

The results from the study examining various ML techniques could accurately classify participants into either high or low performer categories. The algorithms used the percentage of correct responses and reaction times from the VRST to predict participant performance based on throughput scores. Throughput scores consider speed accuracy tradeoffs that many participants make when performing timed assessments (Thorne, 2006) While results varied slightly when classifying performance based on safe zone performance, ambush zone performance, or overall performance, NB classifiers tended to perform well, with around 80% correct classifications. Additionally, many of the SVM algorithms tended to outperform the kNN algorithms. Several of the algorithms which relied on gaussian distributions tended to perform well. However,

classifiers that relied on too few or too many examples for classification rules and cutoffs lead to poorer classification accuracy, likely because of the bias/variance tradeoff. These classifiers included fine gaussian kNN, coarse gaussian kNN, or fine gaussian SVM algorithms tended to lead to poorer classification accuracy.

4.2 VRST Factor Analysis

4.2.1 ANAM

The ANAM had enough evidence for underlying factors based on KMO scores and Bartlett's test (Howard, 2016). Two factors were extracted from the ANAM again one factor was related to response times and one factor was related to accuracy. However, weaker evidence existed for the inclusion of the second factor related to accuracy. When extracting factors using principal axis factoring the diagonal of the correlation matrix is replaced with squared multiple correlation coefficients, these values are used as estimates of the communalities (i.e., the variance shared between the item and all the factors; Tabachnick & Fidell, 2013). PAF replaces the squared multiple correlation coefficients with the communalities and then reruns the analysis, until the analysis changes little from one iteration to the next. While the initial estimates indicated the eigenvalue was likely above one it was reduced to below one, indicating that the factor is likely not explaining a significant portion of variance in scores. Additionally, the accuracy variables did not load onto this factor as strongly as the VRST. Taken together this may indicate that the ANAM Stroop task is mainly measuring response time to Stroop stimuli. The reaction time factor and accuracy factor were only somewhat correlated, indicating that participants may be choosing to either respond with a focus on speed or accuracy.

Participants may be selecting to respond accurately to the Stroop stimuli or may be responding as quickly as possible. As indicated by Thorne (2006), participants tend to select one

of two strategies when performing timed tests. In the first strategy participants may favor response accuracy leading to lowered response time. The other common strategy involves focusing on response speed often at the expense of response accuracy. Throughput is calculated by summing the number of correct responses then dividing by the total time taken for all responses (Thorne, 2006). Therefore, throughput can be used as a measure of participant performance even when participants select different strategies because it can account for tradeoffs participants make between speed and accuracy (Thorne, 2006).

4.2.2 VRST Combined

When VRST data from both safe zones and ambush zones were analyzed together it was found that there were two factors. One of the factors measured response time and the other factor measured response accuracy. The KMO score indicated that the factors were well defined Bartlett's test of sphericity also provided evidence for their existence (Howard, 2016). The variables included in the analysis clearly loaded onto one factor or the other. An oblique rotation was performed on the extracted factors. Oblique rotations allow the axis angle between extracted factors to change, factors can be correlated and tend to produce clearer factor structures than orthogonal rotations (Osborne, 2015). The weak correlations between the two factors within the exploratory factor analysis indicated that the factors are unlikely to be measuring the same underlying constructs. Further, when correlating the extracted factor scores from participants, the correlation was weak and not statistically significant.

Armstrong and colleagues suggested a single factor structure for the VRST because the VRST conditions were correlated (Armstrong et al., 2013). However, in the Armstrong study, only reaction times from the VRST were included. The current study also included a measure of accuracy; including accuracy may have led to multiple factors being extracted in our study.

4.2.3 VRST Safe Zones vs Ambush Zones

The VRST includes safe zones where little activity occurs within the virtual environment and ambush zones where potentially arousing stimuli are experienced. There are some overlapping brain areas which are important for both emotional processing and for cognitive control (Schweizer et al., 2013). Changes in psychophysiological arousal have been observed during the VRST (Wu & Parsons, 2012), therefore separate factor analyses were conducted for safe zones and ambush zones. Findings were generally similar for both safe zones and ambush zones. In both cases KMO scores and Bartlett's test of sphericity indicated that there were underlying factors. Two factors were extracted for both analyses as well. In both cases the first factor was related to response time and the other factor was related to response accuracy. One difference between the analyses was that the two factors were weakly correlated for the safe zones when the factors were rotated but were uncorrelated within the ambush zones. In both cases variables clearly loaded onto a single factor. The response factor from the safe zones was correlated with the response factor from the ambush zones, the same was found for accuracy factors. However, the correlations were weaker than the correlations between then the combined factor analysis. This indicates that while the factors are similar differences between the two conditions likely exist.

When examining throughput from the VRST differences between safe and ambush zone performance were observed. Participants tended to respond more quickly in the ambush zones but also had reduced accuracy. When participants encounter the arousing stimuli, participants may be experiencing greater levels of arousal. Previous research examining the VRST found increased heart rate, respiration rate, skin conductance level when participants experienced ambush zones (Parsons & Courtney, 2018). The VRST may tap more into affective processing

within ambush zones compared to safe zones. This may lead to participants being more likely rely on response styles focusing on response speed at the expense of accuracy.

4.2.4 Comparison of Stroop Tasks

As previously discussed, similarities and differences exist between the VRST and the ANAM Stroop tasks. Results from analyses of both the VRST and the ANAM Stroop suggested a two-factor solution. Again, these were response time factors and accuracy factors. The correlations between the factor scores from the different Stroop modalities were in the expected directions (e.g., higher scores on the response time factor from the VRST were associated with higher scores on the response time factor from the ANAM). The similarities between the factor analytic results provided evidence for convergent validity (Carlson & Herdman, 2012). The ANAM Stroop and the VRST both present Stroop stimuli one at a time, when participants respond to the stimulus the next stimulus appears. Due to single item presentation participants may be focusing on responding to the item as soon as it is presented. Possibly minimizing interference due to surrounding Stroop stimuli when compared to Stroop tasks that present items concurrently (Periáñez et al., 2021).

Correlations tended to be higher between the safe zone factors and the ANAM than between ambush zone factors and the ANAM. This is possibly due to the inclusion of arousing stimuli. As discussed above. Other differences between the VRST and the ANAM Stroop task also exist. For example, the VRST includes complex interference conditions, the goal of the complex interference condition is to increase cognitive load. However, because the ANAM Stroop does not have a complex interference condition it was not included in the current analysis. Correlations have been previously observed between scores from the ANAM Stroop task and VRST. (Armstrong et al., 2013). However, the VRST is considered a higher dimensional

assessment because the ANAM Stroop only includes simple stimulus presentations for example stimuli are presented in the middle of a blank screen within the ANAM Stroop (Parsons & Duffield, 2020). Other research has also indicated that there may be significant differences between measures utilizing virtual reality and those that do not (Neguţ et al., 2015). Some of these differences may be due to many virtual environments being designed to have greater ecological validity (Parsons et al., 2017). The VRST gives participants more contextual information than many other Stroop tasks, participants ride in a simulated HMMWV along a desert road modeled after middle eastern environments, rather than simply look at a sheet of stimuli or stimuli centered in the middle of a blank screen (Parsons et al., 2013). Similar to differences observed between safe zones and ambush zones within the VRST, the ANAM Stroop is unlikely to assess affective processing. Lower dimensional assessments often focus on more purely cognitive abilities (Nejati et al., 2018). However, as noted by Zelazo (2015) both abilities associated with planning, inhibition, and working memory as well as abilities which are important for emotion, motivation, and immediate vs late delayed gratification are considered important when examining real-world behaviors, which the ANAM Stroop task may not be able to do as well as the VRST a higher dimensional task.

4.3 Discussion Machine Learning Analysis

The VRST was initially designed to utilize VR technologies for the examination of multiple variables including the impact of arousing stimuli and interference complexity on Stroop task performance (Parsons et al., 2013). Computerized and VR assessments can be used to create adaptive assessments. Adaptive systems may enable users to achieve more optimal performance or engagement, possibly through increased likelihood of flow states (Parsons & Courtney, 2011).

The effectiveness of the classification algorithm is an important consideration for an adaptive assessment. Several of the algorithms relied on gaussian distributions and tended to perform well. This may be because normalization of data has been found to improve classifier performance in certain instances (Beunza et al., 2019). Data used in the current study could be considered normally distributed and possibly boosted performance of the gaussian based classifiers compared to other types.

4.3.1 Naïve Bayes Performance

NB classifiers assume each predictor is independentClick or tap here to enter text., but NB classifiers also tend to perform well even when this assumption is violated (Arar & Ayan, 2017). A potential benefit of this aspect of the NB classifiers is that they may be able to use provided data more effectively. For example, in the current study, data was used from 157 participants, however after data cleaning and outlier removal, 71 participants did not have complete data. Specifically, 71 participants did not have full data from the combined analysis, 54 from safe zones, and 50 from ambush zones. Because SVM require complete data these participants were not used for the predictions. However, NB classifiers were able to utilize all available data because the classifier uses maximum likelihood estimation, and the predictors are considered independent. The Gaussian NB and kernel NB may have performed similarly because the data was normally distributed. Kernel NB has less stringent requirements for the predictors, it calculates distribution estimates for each predictor leading it to be more computationally intensive, whereas Gaussian NB assumes that the predictors are normally distributed.

4.3.2 Support Vector Machine Performance

The SVMs classifiers also tended to perform well. While the SVM algorithms require participants to not have any missing data (unlike the NB classifiers), these algorithms overall had

high correct classification rates for the included participants. One of the most common types of SVM classifier is the linear SVM which produces a linear function based on weights applied to predictor scores to create a cutoff to divide participants into high and low performers. This classifier often has scores which are more intuitively interpretable. The linear SVM classifier tended to perform well in all three classification situations. Quadratic and cubic SVM raise the kernel function to a power allowing the classifier to have curves which may improve classification performance. When data from safe and ambush zones was combined the cubic SVM classifier performed better than all other algorithms. However, these functions may reduce interpretability due to transformations of kernels. Medium and coarse Gaussian SVMs also performed well; both classifiers use the Gaussian kernel algorithm. While the fine Gaussian SVM also uses Gaussian kernel, likely accuracy was poor due to the bias/variance tradeoff. Bias indicates how closely the classifier matches the training data, and variance indicates how well classifiers perform when applied to test data or when generalized to new data (Belkin et al., 2019). Likely fine Gaussian SVM made predictions that matched the training data too closely, which may have led to poor performance when examining the test data. A 10 v-fold validation was used in the study, this procedure estimates the classifier's ability to correctly classify new data. This procedure trains each classifier on 90% of the data and tests its classification performance on the remaining 10% of data. The procedure is conducted 10 times, with each procedure a separate 10% of the data, maximizing the amount of training and test data available.

4.3.3 k Nearest Neighbors Machine Performance

Finally, kNN algorithms was also used for classification. As stated above, kNN use a system like voting to classify data. These classifiers used various formulas to determine the closest number of neighbors (k) to determine which group a participant belongs to. These

algorithms tended to perform poorly in the current study. Fine, medium, and coarse kNN algorithms used, 1, 10, and 100 neighbors as the number of neighbors in the algorithm respectively. The coarse kNN classifier performed the worst, but it is possible that even though there were more than 100 participants used in the analysis, more data is needed to effectively utilize such a large k value. More effective k values are generally closer to the square root of the sample size (Gareth et al., 2013). The current study had 157 participants, 90% (~141) were used for training the data the square root of this value is 11.8 and as shown in classifying performance in safe zones, ambush zones, and overall performance medium kNN using a k of 10 outperformed both fine and coarse kNN classifiers. Cosine, cubic, and weighted kNN classifiers all used a k of 10 as well, however, cosine and cubic classifiers used different distance formulas to determine which neighbors were nearest, while weighted kNN uses weights to allow the closest neighbors to have a greater impact on the classification compared to neighbors which are further away. In general, it was observed that weighted kNN and kNNs with different distance formulas performed similarly to the medium kNN classifier. Within a different virtual Stroop kNN algorithms tended to perform worse in another virtual Stroop environment (McMahan et al., 2021).

CHAPTER 5

CONCLUSIONS

5.1 Overview

The work examined several areas of technological advancement that can be applied to the field of psychology. Some of these technologies include advancements in computing, virtual reality (VR), statistical techniques, and recording/measurement devices. The current work implemented some of these techniques and technologies.

5.2 Conclusions and Limitations from Factor Analysis

There were some potential issues with the factor analyses of the VRST and ANAM Stroop tasks. For example, sample size was relatively low. A sample size of at least 100 participants is preferred for a factor analysis, however determining the proper number of participants needed to increase stability of the results is not easily determined (Gaskin & Happell, 2014). First, traditionally researchers have argued for sample size heuristics such as having at least 300 participants or other methods such as respondent to variable ratios of 10:1 or even 30:1 (Gaskin & Happell, 2014; Yong & Pearce, 2013). As discussed by Gaskin and Happell (2014) simulation studies can help determine when sample sizes will lead to stable results. Unfortunately, knowing the communalities is one of the most useful methods for determining appropriate sample size, but communalities cannot be determined unless data has already been investigated either from a previous study or as a post-hoc analysis of the data collected. Additionally, participants were not assessed for previous military experience. Participants with military experience may not respond in a similar manner compared to civilians, military experience may lead to changes in arousal and differences in cognitive and affective load due to training.

In conclusion, the current study provided additional convergent validity between a higher dimensional Stroop task the VRST and a low dimensional Stroop task the ANAM Stroop task. When examining response times and percent of correct responses, the VRST and ANAM Stroop task both produced two factor solutions. The solutions produced a factor related to the response times and a factor related to the percent of correct responses. The within assessment factors tended to be uncorrelated or weakly correlated possibly indicating that participants are choosing to either respond as quickly as possible or to focus on responding accurately. Therefore, future work examining Stroop tasks with the ability to measure single-item responses may want to include throughput scores. Throughput scores can combine both factors into a single item capturing both aspects of speed and accuracy.

5.3 Conclusions and Limitations from Machine Learning Analysis

The analysis examined several machine learning algorithms and hyper-parameters for the classification of participant performance for the creation of an adaptive version of the VRST. The results indicated that certain ML algorithms can successfully separate participants into high and low performance categories based on output data from the VRST. Additionally, this study examined hyper-parameters which can influence the performance of ML algorithms. The next step for creating an adaptive VE would be to create rules which would influence the environment itself. For example, if participants are performing poorly during the ambush zone, the VRST might be able to increase user performance by decreasing the number of arousing stimuli presented to the participant. In contrast, if a participant is near ceiling level performance, the participant could be further challenged by increasing the number of arousing stimuli. Other factors from the VRST could also be manipulated such as length of time participants spend

within each zone depending on performance, such that if participants are performing poorly within a zone, additional time could be given to the participants to adequately respond to stimuli.

The classifiers were based on full datasets, but within an adaptive environment, the classifiers will not have access to the participant's scores from zones which the participant has not encountered yet. Previous work from on the VRST suggests that applications such as transfer learning and active class selection may be able to boost classification performance in these settings (Wu & Parsons 2011). The current study determined whether it was possible to categorize participants as high or low performers at all based on output data from the VRST and that the hyper-parameters of the classifiers themselves influenced classification accuracy. While it was identified that participants can be accurately categorized using some of the ML algorithms, additional work could examine if scores collected mid-way through an assessment can be used to categorize participants as high or low performers. Future work may also examine if additional data collected during the VRST, such as psychophysiological data or number of stimuli encountered, may improve classification accuracy. One other limitation of the study was the use of only two categories high and low performance. Future work should include the addition of more categories to fine tune the user performance. The research conducted shows that various ML techniques, particularly NB classifiers can accurately classify participants into high and low performance groups based on the percentage of correct responses and reaction times to Stroop stimuli. SVM classifiers also tended to perform well but may be at a disadvantage for adaptive assessments because they require cases without missing data. Data would likely be readily available for measures such as reaction times or number of correct responses. However, more direct measures such as arousal including EEG or heart rate noisy input data could potentially impact classifier performance. Lastly, kNN algorithms tended to perform the worst.

Additional work is needed for the VRST to be a successful adaptive VE. Further studies may find that when data is collected part way through an assessment performance, accuracies for the classifier's changes.

5.4 General Conclusions

First, a VR based Stroop task, the Virtual Reality Stroop Task (VRST; high mobility multipurpose wheeled vehicle; HMMWV version), was compared to a computerized Stroop task (i.e., from the Automated Neuropsychological Assessment Metrics; ANAM). It was found that both the VRST and ANAM Stroop tasks produced two-factor solutions. For both assessments the first factor related to correct responding and the second factor related to response speed. It was found that these factors were not highly related to each other indicating that participants may be focusing on either responding accurately to stimuli or responding swiftly to stimuli. Further, from the VRST safe zone performance and ambush zone performance were also investigated, while both had similar factor structures are the combined analysis, differences in performance were observed when arousing stimuli were included. The ANAM Stroop tasks provide participants with a limited experience: the assessment tends to be less representative of real-world environments. The ANAM Stroop task and similar measures may be less engaging and provide less diverse ways of presenting stimuli, when compared to VR assessments (Gerjets et al., 2014).

Additionally, machine learning (ML) algorithms for participant performance classification were examined. The study found that some of the ML algorithms were able to accurately predict high or low performance based on correct responses and reaction times from the VRST. The study found that results varied a bit throughout different sections of the assessment, but SVM classifiers tended to perform the best in the current analysis. SVM

classifiers tended to have high correct classifications and AUCs. Also, NB classifiers performed well in the current study but kNN algorithms did not seem to perform as well. When utilizing ML algorithms, it is important to consider their performance and which aspects of the algorithm (i.e., hyper-parameters) may produce the best results (Luo, 2016).

The field of psychology can potentially benefit from advancements in technology. Some of these benefits include improved ecological validity (i.e., the degree to which measures match real-world situations or predict real-world behaviors), improved control over stimuli, ability to collect data from situations which may otherwise be impossible, and allow for additional ways measures can assess and interact with participants. Psychometric properties of newly created high-dimensional VR assessments should be assessed in a manner similar to the creation of new low dimensional assessments. Finally, when evaluating the ability of ML algorithms for classification and potential use for automated assessment, the specific classifier used and hyper-parameters of the classifiers should be examined as they can impact classifier performance.

REFERENCES

Arar, Ö. F., & Ayan, K. (2017). A feature dependent Naive Bayes approach and its application to the software defect prediction problem. *Applied Soft Computing, 59*, 197-209.

Armstrong, C. M., Reger, G. M., Edwards, J., Rizzo, A. A., Courtney, C. G., & Parsons, T. D. (2013). Validity of the Virtual Reality Stroop Task (VRST) in active duty military. *Journal of Clinical and Experimental Neuropsychology*, *35*(2), 113–123.

Badesa, F. J., Morales, R., Garcia-Aracil, N., Sabater, J. M., Casals, A., & Zollo, L. (2014). Auto-adaptive robot-aided therapy using machine learning techniques. *Computer methods and programs in biomedicine, 116*(2), 123-130.

Barua, S., Ahmed, M. U., & Begum, S. (2020). Towards intelligent data analytics: A case study in driver cognitive load classification. *Brain sciences, 10*(8), 526.

Baumeister, R. F. (2016). Charting the future of social psychology on stormy seas: Winners, losers, and recommendations. *Journal of Experimental Social Psychology, 66*, 153–158.

Belkin, M., Hsu, D., Ma, S., & Mandal, S. (2019). Reconciling modern machine-learning practice and the classical bias–variance trade-off. *Proceedings of the National Academy of Sciences, 116*(32), 15849-15854.

Beunza, J. J., Puertas, E., García-Ovejero, E., Villalba, G., Condes, E., Koleva, G., Hurtado, C., & Landecho, M. F. (2019). Comparison of machine learning algorithms for clinical event prediction (risk of coronary heart disease). *Journal of Biomedical Informatics*, *97*, 103257.

Bhavsar, H., & Panchal, M. H. (2012). A review on support vector machine for data classification. *International Journal of Advanced Research in Computer Engineering & Technology (IJARCET), 1*(10), 185–189.

Brunetti, R., Indraccolo, A., Del Gatto, C., Farina, B., Imperatori, C., Fontana, E., ... & Adenzato, M. (2021). eStroop: Implementation, Standardization, and Systematic Comparison of a New Voice-Key Version of the Traditional Stroop Task. *Frontiers in Psychology, 12*, 2041.

Cai, X., & Padoa-Schioppa, C. (2012). Neuronal encoding of subjective value in dorsal and ventral anterior cingulate cortex. *Journal of Neuroscience, 32*(11), 3791–3808.

Carlson, K. D., & Herdman, A. O. (2012). Understanding the impact of convergent validity on research results. *Organizational Research Methods*, *15*(1), 17–32.

Chaby, L. E., Sheriff, M. J., Hirrlinger, A. M., & Braithwaite, V. A. (2015). Can we understand how developmental stress enhances performance under future threat with the Yerkes-Dodson law? *Communicative & Integrative Biology, 8*(3), e1029689.

Chirico, A., Serino, S., Cipresso, P., Gaggioli, A., & Riva, G. (2015). When music “flows”. State and trait in musical performance, composition and listening: a systematic review. *Frontiers in Psychology, 6*, 906.

Cieslik, E. C., Mueller, V. I., Eickhoff, C. R., Langner, R., & Eickhoff, S. B. (2015). Three key regions for supervisory attentional control: evidence from neuroimaging meta-analyses. *Neuroscience & biobehavioral reviews, 48*, 22-34.

Davidson, D. J., Zacks, R. T., & Williams, C. C. (2003). Stroop interference, practice, and aging. *Aging, Neuropsychology, and Cognition*, *10*(2), 85–98.

Diamond, N. B., & Levine, B. (2018). *The prefrontal cortex and human memory.*

Diemer, J., Alpers, G. W., Peperkorn, H. M., Shiban, Y., & Mühlberger, A. (2015). The impact of perception and presence on emotional reactions: a review of research in virtual reality. *Frontiers in Psychology, 6*, 26.

Drey, T., Jansen, P., Fischbach, F., Frommel, J., & Rukzio, E. (2020). Towards progress assessment for adaptive hints in educational virtual reality games. *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*, 1–9.

Evans, J. S. B. (2008). Dual-processing accounts of reasoning, judgment, and social cognition. *Annu. Rev. Psychol.*, *59*, 255–278.

Feroz, F. S., Leicht, G., Rauh, J., & Mulert, C. (2019). The time course of dorsal and rostral-ventral anterior cingulate cortex activity in the emotional Stroop experiment reveals valence and arousal aberrant modulation in patients with schizophrenia. *Brain topography*, *32*(1), 161-177.

Flach, P. A. (2016). ROC analysis. In Encyclopedia of machine learning and data mining (pp. 1-8). *Springer*.

Galatzer-Levy, I. R., Ma, S., Statnikov, A., Yehuda, R., & Shalev, A. Y. (2017). Utilization of machine learning for prediction of post-traumatic stress: a re-examination of cortisol in the prediction and pathways to non-remitting PTSD. Translational psychiatry, 7(3), e1070-e1070.Gareth, J., Daniela, W., Trevor, H., & Robert, T. (2013). *An introduction to statistical learning: with applications in R*. Spinger.

Gaskin, C. J., & Happell, B. (2014). On exploratory factor analysis: A review of recent evidence, an assessment of current practice, and recommendations for future use. *International Journal of Nursing Studies*, *51*(3), 511–521.

Gerjets, P., Walter, C., Rosenstiel, W., Bogdan, M., & Zander, T. O. (2014). Cognitive state monitoring and the design of adaptive instruction in digital environments: lessons learned from cognitive workload assessment using a passive brain-computer interface approach. *Frontiers in Neuroscience*, *8*, 385.

Gross, J. J. (2015). Emotion regulation: Current status and future prospects. *Psychological Inquiry*, *26*(1), 1–26.

Hautamäki, V., Cherednichenko, S., Kärkkäinen, I., Kinnunen, T., & Fränti, P. (2005). Improving k-means by outlier removal. *Scandinavian Conference on Image Analysis*, 978–987.

Heidlmayr, K., Kihlstedt, M., & Isel, F. (2020). A review on the electroencephalography markers of Stroop executive control processes. *Brain and Cognition*, *146*, 105637.

Hossin, M., & Sulaiman, M. N. (2015). A review on evaluation metrics for data classification evaluations. *International Journal of Data Mining & Knowledge Management Process*, *5*(2), 1.

Howard, M. C. (2016). A review of exploratory factor analysis decisions and overview of current practices: What we are doing and how can we improve? *International Journal of Human-Computer Interaction*, *32*(1), 51–62.

Ippolito, M., Ferguson, J., & Jenson, F. (2021). Improving facies prediction by combining supervised and unsupervised learning methods. *Journal of Petroleum Science and Engineering*, *200*, 108300.

Jiao, Y., & Du, P. (2016). Performance measures in evaluating machine learning based bioinformatics predictors for classifications. *Quantitative Biology*, *4*(4), 320–330.

Jin, J., & Maren, S. (2015). Prefrontal-hippocampal interactions in memory and emotion. *Frontiers in Systems Neuroscience*, *9*, 170.

Johannessen, E., Szulewski, A., Radulovic, N., White, M., Braund, H., Howes, D., Rodenburg, D., & Davies, C. (2020). Psychophysiologic measures of cognitive load in physician team leaders during trauma resuscitation. *Computers in Human Behavior*, *111*, 106393.

Kessels, R. P. C. (2019). Improving precision in neuropsychological assessment: Bridging the gap between classic paper-and-pencil tests and paradigms from cognitive neuroscience. *The Clinical Neuropsychologist*, *33*(2), 357–368.

Laeng, B., Ørbo, M., Holmlund, T., & Miozzo, M. (2011). Pupillary stroop effects. *Cognitive Processing*, *12*(1), 13–21.

Lifshitz, M., Bonn, N. A., Fischer, A., Kashem, I. F., & Raz, A. (2013). Using suggestion to modulate automatic processes: From Stroop to McGurk and beyond. *Cortex, 49*(2), 463-473.

Luo, G. (2016). A review of automatic selection methods for machine learning algorithms and hyper-parameter values. *Network Modeling Analysis in Health Informatics and Bioinformatics*, *5*(1), 1–16.

MacLeod, C. M. (1991). Half a century of research on the Stroop effect: an integrative review. *Psychological Bulletin*, *109*(2), 163.

Mahesh, B. (2020). Machine learning algorithms-a review. *International Journal of Science and Research (IJSR).[Internet]*, *9*, 381–386.

Mandrekar, J. N. (2010). Receiver operating characteristic curve in diagnostic test assessment. *Journal of Thoracic Oncology*, *5*(9), 1315–1316.

McCabe, D. P., Roediger III, H. L., McDaniel, M. A., Balota, D. A., & Hambrick, D. Z. (2010). The relationship between working memory capacity and executive functioning: evidence for a common executive attention construct. *Neuropsychology*, *24*(2), 222.

McMahan, T., Duffield, T., & Parsons, T. D. (2021). Feasibility Study to Identify Machine Learning Predictors for a Virtual School Environment: Virtual Reality Stroop Task. Front. *Virtual Real. 2: 673191. Doi: 10.3389/Frvir*.

McMahan, T., & Parsons, T. D. (2020). Adaptive Virtual Environments using Machine Learning and Artificial Intelligence. *ANNUAL REVIEW OF CYBERTHERAPY AND TELEMEDICINE 2020*, 141.

Meyers, J. E., & Vincent, A. S. (2020). Automated neuropsychological assessment metrics (v4) military battery: military normative data. *Military Medicine*, *185*(9-10), e1706-e1721.

Mitra, J., Shen, K., Ghose, S., Bourgeat, P., Fripp, J., Salvado, O., Pannek, K., Taylor, D. J., Mathias, J. L., & Rose, S. (2016). Statistical machine learning to identify traumatic brain injury (TBI) from structural disconnections of white matter networks. *NeuroImage*, *129*, 247–259.

Mosavi, A., Ozturk, P., & Chau, K. (2018). Flood prediction using machine learning models: Literature review. *Water*, *10*(11), 1536.

Nakamura, J., & Csikszentmihalyi, M. (2014). The concept of flow. In *Flow and the foundations of positive psychology* (pp. 239–263). Springer.

Neguţ, A., Matu, S.-A., Sava, F. A., & David, D. (2015). Convergent validity of virtual reality neurocognitive assessment: a meta-analytic approach. *Transylvanian Journal of Psychology*, *16*(1).

Nejati, V., Salehinejad, M. A., & Nitsche, M. A. (2018). Interaction of the left dorsolateral prefrontal cortex (l-DLPFC) and right orbitofrontal cortex (OFC) in hot and cold executive functions: Evidence from transcranial direct current stimulation (tDCS). *Neuroscience*, *369*, 109–123.

Ngure, J. N., Kihoro, J. M., & Waititu, A. (2015). Principal component and principal axis factoring of factors associated with high population in urban areas: a case study of Juja and Thika, Kenya. *American Journal of Theoretical and Applied Statistics*, *4*(4), 258.

Noble, W. S. (2006). What is a support vector machine? *Nature Biotechnology*, *24*(12), 1565–1567.

Norsworthy, C., Gorczynski, P., & Jackson, S. A. (2017). A systematic review of flow training on flow states and performance in elite athletes. *Graduate Journal of Sport, Exercise & Physical Education Research*, *6*(2), 16–28.

Omar, K. S., Mondal, P., Khan, N. S., Rizvi, M. R. K., & Islam, M. N. (2019). A machine learning approach to predict autism spectrum disorder. *2019 International Conference on Electrical, Computer and Communication Engineering (ECCE)*, 1–6.

Osborne, J. W. (2015). What is rotating in exploratory factor analysis? *Practical Assessment, Research, and Evaluation*, *20*(1), 2.

Pan, X., & Hamilton, A. F. de C. (2018). Why and how to use virtual reality to study human social interaction: The challenges of exploring a new research landscape. *British Journal of Psychology*, *109*(3), 395–417.

Parsons, T. D. (2015). Virtual reality for enhanced ecological validity and experimental control in the clinical, affective and social neurosciences. *Frontiers in human neuroscience, 9*, 660.

Parsons, T. D., & Barnett, M. (2019). Virtual Apartment-Based Stroop for assessing distractor inhibition in healthy aging. *Applied Neuropsychology: Adult*, *26*(2), 144–154.

Parsons, T. D., & Barnett, M. D. (2018). Virtual apartment stroop task: Comparison with computerized and traditional stroop tasks. *Journal of Neuroscience Methods*, *309*, 35–40.

Parsons, T. D. & Courtney, C. G. (2018). Interactions between threat and executive control in a virtual reality stroop task. *IEEE Transactions on Affective Computing*, *9*(1), 66–75.

Parsons, T. D., Courtney, C. G., & Dawson, M. E. (2013). Virtual reality Stroop task for assessment of supervisory attentional processing. *Journal of Clinical and Experimental Neuropsychology*, *35*(8), 812–826.

Parsons, T. D. & Duffield, T. (2019). National Institutes of Health initiatives for advancing scientific developments in clinical neuropsychology. *The Clinical Neuropsychologist*, *33*(2), 246–270.

Parsons, T. D. & Duffield, T. (2020). Paradigm shift toward digital neuropsychology and high-dimensional neuropsychological assessments. *Journal of Medical Internet Research*, *22*(12), e23777.

Parsons, T. D., Gaggioli, A., & Riva, G. (2017). Virtual reality for research in social neuroscience. *Brain Sciences*, *7*(4), 42.

Parsons, T. D., Gaggioli, A., & Riva, G. (2020). Extended reality for the clinical, affective, and social neurosciences. *Brain Sciences*, *10*(12), 922.

Parsons, T.D., McMahan, T., & Parberry, I. (2022). Classification of Video Game Player Experience Using Consumer-Grade Electroencephalography. *IEEE Transactions on Affective Computing, 13*(1), 315.

Parsons, T. D., & Reinebold, J. (2011, November). Neuroscience and simulation int erface for adaptive assessment in serious games. In *2011 IEEE International Games Innovation Conference (IGIC)* (pp. 93-96). IEEE.

Parsons, T.D, & Reinebold, J. (2012). Adaptive Virtual Environments for Neuropsychological Assessment in Serious Games. *IEEE Transactions on Consumer Electronics, 58,* 197-204.

Pennycook, G. (2017). A Perspective on the Theoretical Foundation of Dual Process Models. *In Dual Process Theory 2.0* (pp. 5-27). Routledge.

Periáñez, J. A., Lubrini, G., García-Gutiérrez, A., & Ríos-Lago, M. (2021). Construct validity of the stroop color-word test: influence of speed of visual search, verbal fluency, working memory, cognitive flexibility, and conflict monitoring. *Archives of Clinical Neuropsychology*, *36*(1), 99–111.

Plass, J. L., & Kalyuga, S. (2019). Four ways of considering emotion in cognitive load theory. *Educational Psychology Review, 31*(2), 339-359.

Rabin, L. A., Spadaccini, A. T., Brodale, D. L., Grant, K. S., Elbulok-Charcape, M. M., & Barr, W. B. (2014). Utilization rates of computerized tests and test batteries among clinical neuropsychologists in the United States and Canada. *Professional Psychology: Research and Practice, 45*(5), 368.

Reeves, D. L., Winter, K. P., Bleiberg, J., & Kane, R. L. (2007). ANAM® Genogram: Historical perspectives, description, and current endeavors. *Archives of Clinical Neuropsychology*, (22), 15-37.

Rodríguez-Ardura, I., & Meseguer-Artola, A. (2016). E-learning continuance: The impact of interactivity and the mediating role of imagery, presence and flow. *Information & Management, 53*(4), 504–516.

Romine, W. L., Schroeder, N. L., Graft, J., Yang, F., Sadeghi, R., Zabihimayvan, M., ... & Banerjee, T. (2020). Using machine learning to train a wearable device for measuring students' cognitive load during problem-solving activities based on electrodermal activity, body temperature, and heart rate: development of a cognitive load tracker for both personal and classroom use. *Sensors, 20*(17), 4833.

Rozenek, E. B., Gorska, M., Wilczynska, K., & Waszkiewicz, N. (2019). In search of optimal psychoactivation: stimulants as cognitive performance enhancers/U potrazi za optimalnom psihoaktivacijom--stimulansi kao pojacivaci kognitivne funkcije. *Archives of Industrial Hygiene and Toxicology, 70*(3), 150+.

Ruff, C. C., Woodward, T. S., Laurens, K. R., & Liddle, P. F. (2001). The role of the anterior cingulate cortex in conflict processing: Evidence from reverse Stroop interference. *Neuroimage, 14*(5), 1150-1158

Scarpina, F., & Tagini, S. (2017). The stroop color and word test. *Frontiers in psychology, 8*, 557.

Schilbach, L. (2015). Eye to eye, face to face and brain to brain: novel approaches to study the behavioral dynamics and neural mechanisms of social interactions. *Current Opinion in Behavioral Sciences*, *3*, 130–135.

Schweizer, S., Grahn, J., Hampshire, A., Mobbs, D., & Dalgleish, T. (2013). Training the emotional brain: improving affective control through emotional working memory training. *Journal of Neuroscience*, *33*(12), 5301–5311.

Shenhav, A., Cohen, J. D., & Botvinick, M. M. (2016). Dorsal anterior cingulate cortex and the value of control. *Nature Neuroscience*, *19*(10), 1286–1291.

Tabachnick, B. G., & Fidell, L. S. (2013). *Using multivariate statistics (6th ed.)*. Pearson Boston, MA.

Tarnanas, I., Schlee, W., Tsolaki, M., Müri, R., Mosimann, U., & Nef, T. (2013). Ecological validity of virtual reality daily living activities screening for early dementia: longitudinal study. *JMIR Serious Games*, *1*(1), e2778.

Thorne, D. R. (2006). Throughput: a simple performance index with desirable characteristics. *Behavior Research Methods*, *38*(4), 569.

Vural, M. S., & Gök, M. (2017). Criminal prediction using Naive Bayes theory. *Neural Computing and Applications, 28*(9), 2581-2592.

Wekselblatt, J. B., & Niell, C. M. (2015). Behavioral state—getting “in the zone”. *Neuron, 87*(1), 7-9.

Woodhouse, J., Heyanka, D. J., Scott, J., Vincent, A., Roebuck-Spencer, T., Domboski-Davidson, K., O’Mahar, K., & Adams, R. (2013). Efficacy of the ANAM General Neuropsychological Screening Battery (ANAM GNS) for detecting neurocognitive impairment in a mixed clinical sample. *The Clinical Neuropsychologist*, *27*(3), 376–385.

Wu, D., Courtney, C., Lance, B., Narayanan, S.S., Dawson, M., Oie, K., & Parsons, T.D. (2010). Optimal Arousal Identification and Classification for Affective Computing: Virtual Reality Stroop Task. *IEEE Transactions on Affective Computing, 1*, 109-118.

Wu, D., Lance, B., & Parsons, T.D. (2013). Collaborative Filtering for Brain-Computer Interaction Using Transfer Learning and Active Class Selection. PLOS ONE, 1-18.

Wu, D., & Parsons, T.D. (2011). Active Learning for Arousal Classification. *Lecture Notes in Computer Science, 6975*, 132-141.

Wu, D., & Parsons, T.D. (2012). Customized Cognitive State Recognition Using Minimal User-Specific Data. *Proceedings of the Military Health Systems Research Symposium*, Fort Lauderdale, FL, August 2012

Wu, D., & Parsons, T.D. (2011). Inductive Transfer Learning for Handling Individual Differences in Affective Computing. Lecture Notes in Computer Science, 6975, 142-151.

Yong, A. G., & Pearce, S. (2013). A beginner's guide to factor analysis: Focusing on exploratory factor analysis. *Tutorials in Quantitative Methods for Psychology*, *9*(2), 79–94.

Zahabi, M., & Abdul Razak, A. M. (2020). Adaptive virtual reality-based training: a systematic literature review and framework. *Virtual Reality*, *24*(4), 725–752.

Zelazo, P. D. (2015). Executive function: Reflection, iterative reprocessing, complexity, and the developing brain. *Developmental Review*, *38*, 55–68.