3D-Trail-Making Test: A Touch-Tablet Cognitive Test to Support Intelligent Behavioral Recognition

Raniero Lara-Garduno* Texas A&M University
Takeo Igarashi† The University of Tokyo
Tracy Hammond‡ Texas A&M University

ABSTRACT

The decades-old Trail-Making Test has established itself as an effective and versatile cognitive testing tool. However, its reliance on a paper-and-pencil method of administration severely limits its capabilities for quantitative evaluation. We evaluate input data via sketch recognition algorithms on a digitized version of the Trail-Making Test, as well as exploring the viability of a novel touch-based version designed around the basic concepts of the traditional paper-and-pencil exam. Two quantitative studies helped evaluate the viability of these digital examinations: the first builds a normative data set that shows a stable increase in test completion times with a lower overall skewness, while the second indicates the new examination's capability to perform behavioral classification based on participant nationality by correctly classifying input behavior with up to 96% accuracy.

Keywords: cognitive test, behavioral recognition, touch

Index Terms: Human-computer interaction—Interaction Devices—Touch Screens; Human-computer interaction—Interaction Paradigms—Graphical User Interfaces; Human-computer interaction—Interaction Design—Activity Centered Design

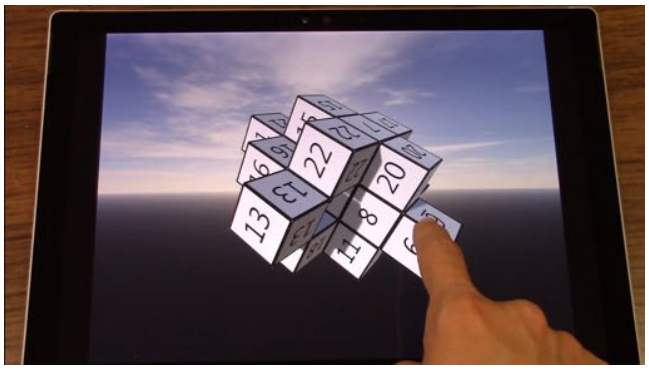

Figure 1: A participant completing the A variation of the 3D-Trail-Making Test.

1 INTRODUCTION

1.1 Description of the Trail-Making Test

The Trail-Making test (TMT) is a set of connect-the-dots exercises completed on paper and pencil. 25 labeled dots are printed onto a piece of paper and the participant is required to connect the dots in the correct order without lifting their pen. The correct order depends on which of the two variants the participant is completing; in the "A" variant, the correct order are dots labeled "1" to "2" to "3" to "4", etc., while the B variant has the order alternating between numbers and letters, "1" to "A" to "2" to "B". The tests are typically administered in pairs, with A immediately followed by B. Variant B is considered more difficult to complete even among cognitively healthy participants [31]. Dots are placed to necessitate searching, as the next dot in the sequence is frequently not in the immediate vicinity of the previous one. The TMT was designed to assess attention, speed, and mental flexibility, and has been actively used in clinical neuropsychology for various decades [29].

1.2 Uses in Clinical Neuropsychology

Neuropsychologists employ a variety of cognitive examinations through which to observe a patient's behavior. Some of these involve manipulating physical objects, answering a series of oral questions, solving exercises that integrate drawing exercises. These tests typically involve paper and pencil, and the patient is required to complete an examination by sketching, drawing on objects, crossing out or circling shapes, and other forms of interaction using pen on paper.

The TMT was conceived and originally introduced as a testing tool in the field of clinical neuropsychology to assess general intelligence. Shortly thereafter, clinicians found utility in the test's ability to aid in the detection of cognitive abnormalities. Clinicians who administer the Trail-Making test observe the patient's behavior according to a predetermined checklist. Among the items include sitting posture, how well they maintain eye contact, reaction to mistakes, and their efforts in remembering the next item in the sequence among others [29]. Tests produce one quantitative performance metric in the form of the time taken to complete the test, rounded to the nearest second. A clinician compares the score against established normative data depending on which age classification the patient belongs to. The lengthy diagnosis process combined with the increased emphasis for early detection due to the climbing proliferation of Alzheimer's disease and other types of dementia[1] has maintained the TMT's relevance in modern diagnoses.

1.3 Uses in Other Fields

One of the landmark traits of the TMT is in its reliable utility as a general tool to test cognitive function. It has been shown to be sensitive to a wide variety of differing behaviors. It is used in sports to assess the extent of effects from mild head injuries [1]. These tests have shown to be an improvement over a subjective report of mild traumatic brain injury symptoms following injuries in sports [9]. The TMT has been shown to be sensitive to both age and education among healthy participants in various populations such as Japan [15], Brazil [14], and Portgual [4], exemplifying its versatility across cultures and geographic regions. It is used to gauge the effects of drug addiction such as cocaine and alcohol on cognitive functions among participants [12]. It is also sensitive to sleep deprivation such as when studying sleep apnea/hypopnea [5]. The TMT also is used in the military, with extensive studies being done to gauge the effect

*e-mail: raniero@tamu.edu
†e-mail: takeo@acm.org
‡e-mail: hammond@tamu.edu

[1]Centers for Disease Control and Prevention: www.cdc.gov/features/alzheimers-disease-deaths/index.html

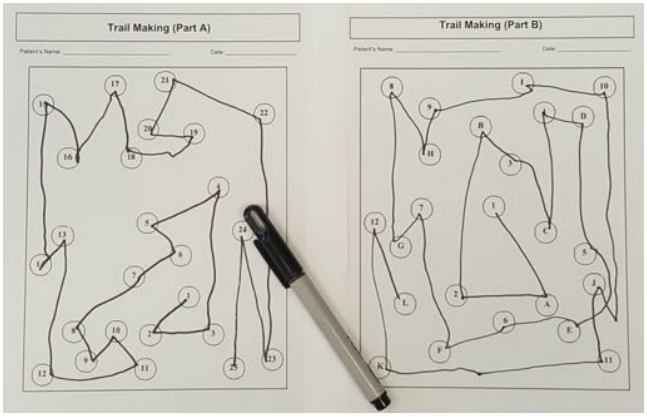

Figure 2: Completed paper-based Trail-Making tests, both A and B variations.

of PTSD on combat veterans [19]. In short, the TMT is sensitive to changes in behavior based on brain injuries, traumatic events, sleep deprivation, age, education, drug addiction, cultural background, and geographic location. This has allowed the TMT's expansion of utility across a wide range of purposes over the decades of its existence.

1.4 Challenges of Assessing TMT Participants

The TMT's sole quantitative measure is the participant's completion time, rounded to the nearest second [31]. While this facilitates analysis of participant performance, it also discards large amounts of context-sensitive data that may yield insight into the participants' state of mind. Whether the administrator is attempting to identify brain injury, mild cognitive impairment, cultural or education differences, PTSD or inebriation, much of the assessment is left to subjective evaluation. The only quantitative comparison that can be made is to whether their score is within the range of "normal" scores according to established normative data. As mentioned above, the test's primary use in clinical neuropsychology still relies on external factors such as sitting posture and reactions to mistakes to gauge a patient's cognitive state, resulting in a lengthy testing process.

The TMT also does not test a participant's perception of objects in a 3-dimensional space and any cognitive functions related to 3-dimensions such as image recognition of rotated objects, object permanence, and depth perception. Since the TMT's introduction in the 1940s, newer research into human cognition has more recently identified general perception of 3-D space as an important indicator in a participant's cognitive state [30].

1.5 Proposed Solution

The advancements of consumer-level computing devices can help improve the TMT in two ways: **(1)** analyze the direct input data using machine learning to produce more quantitative behavioral recognition, and **(2)** support the development of novel variations that allow testing for new cognitive functions not otherwise possible on the paper-and-pencil format. We explore both potential improvements by digitizing the TMT and developing a new touch-based test that specifically integrates concepts related to 3-dimensions into the existing TMT.

The novel test, called 3D-Trail-Making Test (3D-TMT) consists of a simplified input modality and a high input sampling rate designed to collect large volumes of usage data in the short period of time typically required to complete this test. This version implements design decisions specifically aimed to expand on the weaknesses inherent to paper-based tests, such as the collection of quantitative data that can be linked to behavioral patterns, and the testing of a participant's visuoperceptual capabilities through the test's required perception of a dynamic virtual 3-dimensional object. By collecting rich amounts of usage metrics as participants complete our examination, we can classify participants based on differences in behavioral patterns informed by a participant's nationality. This shows that like the original TMT, the 3D-TMT appears sensitive to nationality, suggesting the examination's flexibility as a behavioral assessment tool with the added benefit of collecting substantial amounts of usage data for quantitative analysis.

2 Related Work

The field of Human-Computer Interaction maintains an interest in neuropsychological assessments due to the near-direct mapping of a participant's drawings with their cognitive state. In this area of study, HCI researchers have leveraged recent advancements in consumer-level technology for two primary purposes: to facilitate analysis of existing cognitive tests, and to explore the possibilities of novel assessments.

2.1 Leveraging Technology for Existing Cognitive Tests

Researchers have previously used segmentation and recognition techniques of drawn shapes for the purposes of assessing cognitive testing performance. Souillard-Mandar et. al developed machine learning algorithms to detect behavioral abnormalities on a digital version of the Clock Drawing Test [28]. Moetesum et. al uses distorted shape recognition to help automatically grade the widely used Bender Gestalt Test (BGT), a drawing test similar in style as the TMT [22, 23], but does not share the TMT's same versatility or ubiquity. Nazar et. al advanced the computerized grading process of the BGT by employing convolutional neural networks [24]. Similar techniques have been employed for the Rey-Osterrieth Complex Figure Test, which consists of copying a complex figure consisting of various simple geometric shapes arranged together. Canham et. al performed some early work in identifying only some individual shapes of the larger complex figure [3], while more recent work from the same author further developed recognition of structures from the naturally distorted Rey-Osterrieth figure drawn by a human [2].

These works establish the concept of leveraging modern machine learning techniques to recognize behavior from neuropsychological exams, but have not worked with the TMT previously. We have previously been able to perform these sketch recognition technologies on a digital version of the TMT to correctly classify a participant's age based on their digital pen input [20].

2.2 Leveraging Technology for Novel Cognitive Tests

A natural extension of this field of research is in the development of novel cognitive examinations. For instance, the research of Zham et al. utilizes variations in pen pressure as a patient draws a spiral as a way to identify different stages of Parkinson's disease [35]. Drotár et al. applies additional handwriting kinematics metrics across large amounts of samples of a participant's handwriting [8].

Researchers have also attempted to use data collection methods in other activities. Jiang et al. utilized virtual reality to analyze the state of a user's wayfinding abilities, particularly observing users with cognitively degenerative conditions [16]. Zavala-Ibarra et al. proposed an architecture for monitoring older patients through the use of ambient video games in order to monitor their health [34]. Drew et. al took a more active approach in aiming to improve perceptual motor skills in elderly populations through their participation in arcade-type video games [7]. Jimison et. al tasked participants with playing 9 different computer games to assess different levels of cognitive function, aiming to improve the identification of behavioral trends in participants [17]. Gong et al. integrated 3D-vision glasses to extract behavioral features and demonstrates that extracted information can be used for cognitive health monitoring by providing detailed physiological information [13]. These research

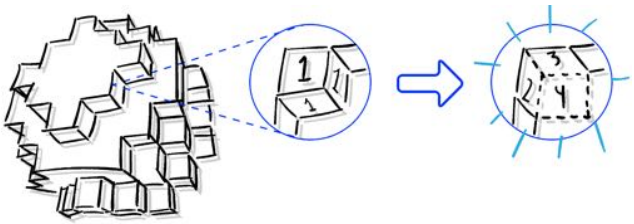

Figure 3: A design drawing depicting the test sphere, made of labeled boxes. Participants tap with their finger on the labeled boxes in the correct order, revealing more labeled boxes of higher numbers underneath.

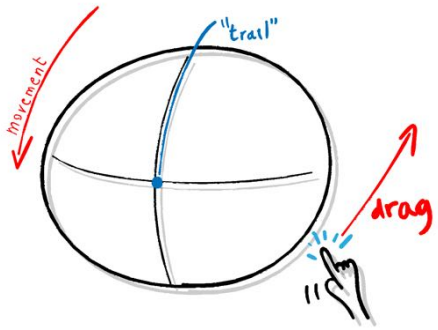

Figure 4: The participant taps and drags the screen to rotate the test sphere. The program samples the sphere's position dozens of times per second, serving as the movement "trail" data to be analyzed for classification.

projects distinctively use principles behind human-computer interaction to identify a diagnosed patient's needs and behavior for assistive technologies.

Our research focuses on using **existing** cognitive examinations to help inform the design of novel tests. The longevity of the TMT is largely due to its ease of use by participants of all ages, cultures, and levels of education, so we believe the design of novel tests must focus on this ease of use and be designed around this ideology. Costly sensors, VR, and AR hardware may prove prohibitively expensive to be used widely, whereas conventional mid and low-end touch tablets and phones continue to rapidly decline in price. Additionally, novel hardware like VR/AR and various sensors may require significant set-up time for test administration. On these complex setups, the tasks themselves are likely to be complex as well, introducing a possible confounder where user performance may decrease in quality due to unfamiliarity with the hardware rather than due to behavioral, cognitive, and cultural differences as is desired. These factors influenced our decision to anchor our design around touch tablet technology.

3 Design

Conceptually, our novel 3D-Trail-Making Test is an intelligent cognitive testing computer application designed to leverage modern mobile capabilities to ultimately facilitate behavioral recognition. Paper-based examinations are used due to the ease of administration and quick completion times, but they lack the inherent ability to test for more complex cognitive functions such as difficulties with depth perception and 3-dimensional spatial relationship of objects. Mendez et al. describes the ways in which neuropsychologists test for these difficulties, which still involves drawing and observing flat images on a paper to simulate 3-dimensional perception [21]. Other forms of testing visuospatial capabilities have included complex experiments, such as the work of Prvulovic et al. which involved custom made foam pads, projectors, frosted screens, and several different stimuli per participant [26]. The complexity of using custom-built objects and the fact that experts already test depth perception from simulating the perception of depth on a flat surface motivated the design of a touch-based mobile application on a tablet.

After careful consideration of the existing technology applications for early diagnosis, we identified five key components for a new examination tool:

- Leverage advantages of existing clinical neuropsychological diagnosis methods by **adapting an existing paper test** into a more modern examination
- **Preserve** the paper test's ease of administration and speed of test completion while testing **new aspects** of a participant's cognitive abilities that a paper test cannot.
- Carefully design the user experience to develop a **streamlined, intuitive interface** that captures a user's **behavior** rather than examine their proficiency with consumer electronics
- **Capture** usage data on a scale appropriate for machine learning classification
- Provide behavioral usage data that can be used to **intelligently** identify differing behaviors among participant groups

3.1 Design Overview

This 3-dimensional test contains both versions of the traditional Trail-Making test, wherein version A has all the boxes labeled in ascending numbers, and version B has the boxes alternating between numbers and letters. Like in traditional 2-dimensional paper-based exams, the correct sequence is different for both the A and B variations and participants complete both exams as a pair. This test was created on the Unity game engine platform and compiled to run on Windows 10.

3.1.1 Input

The classic Trail-Making test serves as the primary basis for this new examination with some differences to better leverage modern tablet technology in the diagnosis process. Rather than placing every labeled dot on a flat 2-dimensional surface with every dot visible to the participants at once, we have rendered every "dot" as interactive 3-dimensional boxes. The boxes are arranged in a spherical pattern and float in an empty virtual space, and with one finger participants drag anywhere on the screen to rotate the sphere on every axis at its center point. The rotation was designed specifically to emulate the rotation gestures of Google Earth [2] due to its familiarity and simplicity. Rotation speed is calculated by directly measuring how much the finger moved across the screen (X and Y, measured in pixels) each frame, multiplied by a constant of 0.5 after testing different rotation speeds in pilot studies. With the same finger, the participant taps on the next box in the sequence. If the correct box is tapped, it disappears with a shrinking animation to clearly indicate correct input. If the incorrect box is tapped, it changes its color to red and remains static to indicate a mistake has been made. When boxes disappear, they reveal more boxes hidden in deeper layers of the sphere. This design is intentional since our goal is to create a dynamic examination wherein not every number or letter is visible at once. This encourages participants to be closely engaged with the exam at all times, since the dynamic nature of the sphere geometry will require their constant re-evaluation of the state of the exam.

[2]Google Earth: google.com/earth/

3.1.2 Ease of Use

The one-finger input modality is intended to mirror the TMT in its ease of use; participants of all age brackets, cultures, educational backgrounds and cognitive states are expected to be able to reasonably engage with the test [11]. When designing our test, the commitment to ergonomic design led us to avoid using Virtual Reality and Augmented Reality despite the more realistic depictions of depth; these relatively cutting-edge technologies are more expensive to deploy than a touch tablet, require more complex setups, and inherently demand a level of comfort with more complex technology. Our intention is to avoid testing a participant's acumen in consumer electronics, and instead to simply produce data directly sensitive to their cognitive state.

3.2 Visual Design

3.2.1 Depiction of Free Space

We have also made efforts to avoid possible confusion resulting from an interactive 3-dimensional object projected into a 2-dimensional screen. We took special consideration in several aesthetic design choices to ensure a participant of any age can clearly differentiate between objects of varying depths and angles. During the test, the sphere is surrounded by a realistic depiction of a sky in the distance in order to clearly indicate that the sphere is floating in a free, empty space and that the sphere's motions will not be interrupted by collision with other objects.

3.2.2 Depiction of Depth

The background depicting a clear afternoon sky also communicates to the user to expect daytime lighting, which we have implemented into the examination with a global illumination system. This heightens the depiction of depth as the sphere of boxes is rotated. The boxes are also capable of casting shadows on each other, with shadow lengths dynamically changing depending on the sphere's location relative to the source of global illumination. This produces very clear distinctions between boxes that differ in depth, as without shadows boxes might seem at the same depth relative to the screen. To further communicate the notion of depth of a 3-dimensional object, each box is rendered with a slight metallic material so that the box subtly shines as it is being rotated. We carefully muted this shine effect so that the reflection does not impede a participant from properly reading the box's number or letter. To make rotation as intuitive as possible, the center box is always the last box of the examination since it is also the center point of rotation of the entire sphere. This serves as a visual anchor for the participant so that rotation is not disorienting even toward the end of the exam when the cluster of boxes no longer forms a spherical shape.

3.3 Replay of Completed Tests

The fine granularity of recorded input data allows us recreate the entirety of participant's exam by reading the file in sequence and advancing the position of the sphere and tapped-box actions with the proper timing and speed. As a result, we have created an additional "Replay" functionality so that a participant or researcher can view the completed test as many times as desired. This functionality also displays a translucent trail of the participant's finger drag across the screen to directly see their input. This "Replay" enhances the observer's ability to produce qualitative analysis, although it is not used for the quantitative evaluation we present in the following section.

3.4 Effects in Cognitive Load

The 3D-TMT deliberately introduces two cognitive tasks not found in the original paper TMT: performing mental rotation tasks, and manipulating a test that dynamically introduces and removes elements. The former is tested by box labels rotating naturally as the participant rotates the test sphere (resulting in boxes that may be skewed, angled, or upside down), while the latter is tested through the mechanic of disappearing cubes that reveal more cubes underneath. These two tasks have been explored in clinical neuropsychology previously, and frequently involve the introduction of new tests [6, 30, 33]. In leveraging the computing power of modern tablet technology, our interest is in carefully designing tests that intuitively combine tasks already known to be explored in the field of clinical neuropsychology. Traditional paper tests requires only one or two of these tasks being tested at one time [32]. As a result, the 3D-TMT is observed to have a higher cognitive load. We recognized this would likely mean a higher average time to test completion than the traditional TMT, a prediction supported by our evaluation results.

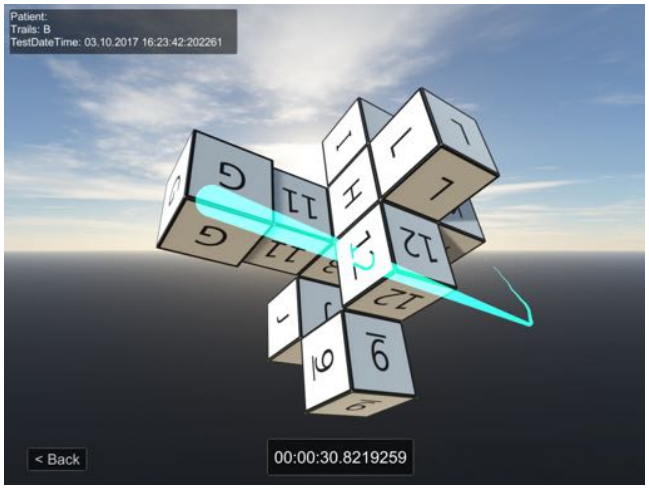

Figure 5: The interface of the Test Replay feature in the 3D-Trail-Making Test application. It provides a real-time playback of a participant's movements. Additional UI features include a running clock, participant information on the top left, and a real-time trail of captured finger movements synchronized with the sphere rotations and finger tap actions.

4 System Evaluation

Cognitive examinations, specifically the TMT, have been previously administered to overseas populations and compared to existing normative data, such as the work of Kim et al. [18], Seo et al. [27], and Cavaco et al. [4]. Our first of the two quantitative studies presented in this paper compares differences between the performance baseline of this examination with that of the traditional paper-and-pencil version. The second quantitative study explores the recognition capabilities of the 3D-TMT, aiming to differentiate between two populations in a way that would be impossible to do with a typical paper-and-pencil examination. All participants in this study are cognitively healthy.

4.1 Normative Studies

4.1.1 Experiment Design

In order to establish a performance baseline for the 3D-TMT, we conducted normative studies with cognitively healthy individuals with a methodology that mirrors that of the existing paper-and-pencil TMT. The 3D-TMT's performance score in this normative study mirrors that of the method of scoring the examination of the traditional TMT, which is "expressed in terms of the time in seconds required for completion of each of the two parts of the test" [29]. For instance, a test with a score of 26 indicates that a participant completed the test in 26 seconds. Existing normative data for the TMT includes dividing the test population into different age groups,

Table 1: Summary of the normative data collected for test variations A and B across paper, digital, and the new 3D examination. Tombaugh's normative data is the first row for reference.

Trail-Making Test A - Completion Time (seconds)

	Mean	Std. Dev	Median	Min-Max	Skewness	Kurtosis
Tombaugh Normative Data	22.93	6.87	21.70	12-57	1.64	4.46
Paper-based TMT	22.57	5.69	21	15-42	1.36	3.24
Digital TMT	23.97	6.61	23	16-46	1.93	4.63
3D-TMT	53.5	15.24	50.5	35-79	0.45	-1.32

Trail-Making Test B - Completion Time (seconds)

	Mean	Std. Dev	Median	Min-Max	Skewness	Kurtosis
Tombaugh Normative Data	48.97	12.69	47.0	29-95	0.91	0.92
Paper-based TMT	42.93	13.80	42.5	27-102	2.73	11.26
Digital TMT	50.86	15.97	47.5	30-99	2.01	4.24
3D-TMT	58.83	15.08	56.5	38-92	0.57	-0.61

one of which is the age group of participants between the ages of 18-24. Normative data can be established with the data of around 30 participants. Following established protocols, our quantitative study involves 30 cognitively healthy participants between the ages of 18-24.

In order to explore the level of consistency between the paper-and-pencil TMT and the 3D-TMT, we conducted a within-subjects study in which each participant in our user study was tasked with completing three sets of Trail-Making tests. The first was a regular paper-based examination administered using the protocol as outlined in Compendium of Neuropsychological Tests. The second set of tests tasks the participant with completing a digital version of the same connect-the-dots exercise by drawing on a Microsoft Surface Pro 4 with a Surface Pen. This simulated the paper-based examination in every way, but with a different layout and positioning of the dots so as to avoid data skew through a Learning Effect. This second test also serves as an anchor point linking the traditional TMT and the touch-based 3D-TMT; establishing consistency in performance between using a paper-based exam and one with a touch-enabled tablet eases concerns about introducing a confounding effect by having to use a tablet for the completion of the 3D-TMT. The third set of tests is the 3D-Trail-Making test completed on the same Microsoft Surface Pro 4 with the participant using his or her finger as input. For each of these three tests, the participants completed both the A variation of the test (order 1, 2, 3, 4) as well as the B variation (order 1, A, 2, B).

4.1.2 Results

Table 1 shows the summary of the collected data from 30 cognitively healthy individuals between the ages of 18-24. For comparison, the established normative data Tombaugh's normative data that is stratified by age [31]. Skewness and kurtosis are two statistical measures typically reported in normative data studies. Skewness is the measure of symmetry where the data distribution is symmetric if it looks the same to the left and right of the center point, while kurtosis is a measure of how heavy the distribution tails are, where high kurtosis values indicate higher presence of outliers[3]. The general performance across the digital version falls in line with that of the paper version, although performance baselines change for the 3D-TMT. The 3D-TMT test A's mean and standard deviations are considerably higher than the either of the two 2-dimensional examinations (paper or digital), and while test B showed a noticeable increase as well in completion time, its increase is considerably less pronounced. Also of note is the fact that the performance across both tests A and B are remarkably close to each other, indicating that across all cognitively healthy participants, expected time to complete the test is more uniform for both tests A and B.

[3]Measures of Skewness and Kurtosis: itl.nist.gov/div898/handbook/eda/section3/eda35b.htm

4.1.3 Discussion

Time to test completion has appreciably changed for both tests A and B for the 3D-TMT. Specifically, test A's completion time has nearly doubled in time. Despite the time increase, we do not consider this detrimental since the average completion time is still under a minute and within the normal range for neuropsychological examinations. Observing a difference in average time to completion alone is not a measure of test efficacy; every neuropsychological test has its own normative time to completion. A more accurate measure of efficacy is consistency among participants of the same classification.

Overall, observed participant behavior remained consistent in nature across all forms of the examination for tests A and B. The changes in test completion time can be attributed to the changes inherent to the different format and increased cognitive load of the examination. It is important to note that sknewness for the 3D-TMT is lower than even that of the skewness presented in Tombaugh's established paper-based normative data, suggesting that test completion times are more consistent between healthy participants. This results in observably consistent behavior among healthy participants, which provides us with strong results as our first set of normative data for the 3D-TMT.

4.2 Classification Capabilities of the 3D-TMT

4.2.1 Experiment Design

The 3-D Trail-Making Test was designed to produce rich volumes of data in order to facilitate machine learning classification. The simplified input modality results in two possible labeled actions: "tap" and "drag". Each of these actions is logged into the usage metrics data along with a time-stamp of when these occurred, the cartesian coordinates of the center box of the sphere, and the x-y coordinates of where the participant's finger was touching the screen. New actions are logged as "Drag" any time the finger is placed and a change in finger position is detected. "Tap" actions get two additional labeled points: which box the participant tapped, and whether it was the next correct one in the sequence. New "Tap" or "Drag" data samples are created only as the participant completes said action. That is, there are no "no action" or "blank" action logs to conserve space, and participant pauses are still preserved by finding the difference between the samples' time-stamps. In total, participants log actions at an average magnitude of several dozen actions per second, typically resulting between 1,000 to 2,000 actions per test. Each usage metric log is saved into a text file along with a participant's anonymized ID number, age, which variant of the exam this log belongs to (A or B), and the date and time that this exam was taken.

This quantitative study is based in concept on the previously mentioned Trail-Making Test studies that analyze populations across different geographic regions [4, 14, 15, 18, 27]. A total of 52 partic-

Table 2: Total list of features used in this classification experiment. "[1-25]" indicates each feature was split into an additional 25, one per the data captured in between each of the boxes. "[X,Y,Z] indicates each feature was further split into 3, one per the direction in which the movement is cumulatively captured.

Behavioral Features
TotalMistakes
TotalTestTime
TimeStartDelay
BoxTimes [1-25]
TimeDragPositive [X,Y,Z] [1-25]
TimeDragNegative [X,Y,Z] [1-25]
DeltaDragPositive [X,Y,Z] [1-25]
DeltaDragNegative [X,Y,Z] [1-25]
AverageTimeDragPositive [X,Y,Z]
AverageTimeDragNegative [X,Y,Z]
AverageDeltaDragPositive [X,Y,Z]
VelocitySamples [X,Y,Z] [1-25]
Magnitudes [1-25]

Table 3: Features yielding the best classification results for the A and B versions of the 3D-Trail-Making Test

Test A	**Test B**
DeltaDragPositiveY_23	BoxTimes_1
DeltaDragPositiveZ_12	BoxTimes_25
DeltaDragPositiveZ_18	TimeDragPositiveZ_2
DeltaDragNegativeX_6	TimeDragPositiveZ_21
DeltaDragNegativeX_10	TimeDragNegativeX_24
DeltaDragNegativeX_16	DeltaDragPositiveY_6
DeltaDragNegativeX_17	DeltaDragPositiveZ_4
DeltaDragNegativeX_25	DeltaDragPositiveZ_11
DeltaDragNegativeY_12	VelocitySamplesZ_7
DeltaDragNegativeZ_13	VelocitySamplesZ_8
DeltaDragNegativeZ_14	VelocitySamplesZ_25
DeltaDragNegativeZ_21	Magnitudes_8
VelocitySamplesZ_5	
VelocitySamplesZ_9	
VelocitySamplesZ_15	
VelocitySamplesZ_16	

ipants completed both variations of test A and B. Of these 52, 32 of them live in the US and identify as having American nationality, and 20 were native citizens of Japan currently residing there. All members of both groups were cognitively healthy and both had an age range of 18-45 years of age.

4.2.2 Feature Calculation

Data features were calculated with the intention to capture participant's behavioral patterns. This is motivated largely due to multiple qualitative observations that we made during the user studies. For instance, members of the Japanese population would more frequently rotate in only one direction as they searched for the next box in the sequence, and they tended to rotate the sphere at a higher speed than that of the American participants. For this reason, our calculated features focused on speed and direction of movement in each of the sphere's three axes. Time to completion was included as well due to the test's relation to the paper-based TMT, as well as timing data in finer granularity. Along those same lines, time spent moving in each direction was also included as a feature. Every feature calculated is shown in Table 2.

"TotalMistakes" is the total amount of mistakes that the participant made over the course of the test. "TimeStartDelay" is the amount of time that a participant took between starting the test and beginning to control the sphere to find the first box. "BoxTimes" is the time taken to correctly tap the next box in the sequence (e.g., the time taken between Box 1 and 2, box 2 and 3, etc.). "TimeDragPositive" is the amount of time that a participant spent rotating the test sphere in the direction that increased the coordinate value (X, Y, and Z were calculated separately). "TimeDragNegative" is, similarly, the amount of time a participant spent rotating in the opposite direction. "DeltaDragPositive" is the distance that the test sphere traveled while the participant rotated the sphere in the positive direction. This was calculated as a separate feature since at different speeds a participant may cover different distances. "DeltaDragNegative" similarly captures distance traveled in the negative direction. The averages of each of these described features across the entire tests were also calculated and included as features. Finally, the velocity vector of the X, Y and Z coordinates is captured in order to track the speed of the participant's movements. The velocity vector is calculated with the formula:

$$v = \frac{[x_t - x_{t-1}, y_t - y_{t-1}, z_t - z_{t-1}]}{timeDiff} \quad (1)$$

"Magnitudes" are the calculated magnitudes of the velocity vector using the formula:

$$v_m = \sqrt{v[x]^2 + v[y]^2 + v[z]^2} \quad (2)$$

Some features were further segmented into each box, indicated by "[1-25]" in Table 2. For example, "TimeDragPositiveX" was divided further into 25 segments, one per box, to indicate the time that the participant spent moving the test sphere in the positive X direction between box 1 and 2, 2 and 3, etc. This would allow us to analyze the participant's behavior on a per-box basis rather than just across the entirety of the test, since we anticipated that patterns might emerge between specific boxes.

This resulted in a total of 440 features, the vast majority of which we anticipated would be culled during feature analysis as we introduced this data into our classifier. We used the standard Naive Bayes classifier, as it is optimal for the Binary classification that we intended to perform between a Japanese and an American population. Classification was supported by 10-fold cross validation.

4.2.3 Results

Classification was performed with the Weka data mining software [10]. Table 4 shows the subset of features across both variations A and B of the 3D-TMT that yielded the highest accuracy in classifying examinations as belonging to either an American or Japanese participant. These features support our earlier intuition that differentiation may be more effective if we focused on tracking the specific direction of motion. For test A, moving the test sphere in the negative X direction across greater distances significantly contributed to the country of origin of the participant. A total of four features related to velocity support the observation that Japanese participants moved the sphere at greater speeds. For test B, a similar result is observed with respect to emphasis of movement velocity, albeit with fewer features needed for classification and the inclusion of timing data for two boxes.

Table 5 shows the detailed accuracy statistics separated by class for both the A and B variation of the 3D-TMT. Specifically, for the A variation, the test was correctly classified as belonging to either an American or Japanese participant with an accuracy of 96.154% and an F-Measure of 0.962. For the B variation, the classification had an accuracy of 88.235% and F-Measure of 0.883.

4.2.4 Discussion

As previously mentioned, "DeltaDrag" refers to the total distance that the participant moved in any particular direction between boxes, split into either a positive or negative direction, and further split

Table 4: A small set of user test data (10 samples) as it is recorded on the text file. The format for each line is: Action Type, X, Y, Z coordinates of the center "anchor" box, X, Y coordinates of the finger touch for replay purposes, Cube ID of tapped box (only if Action Type is "Tap"), whether the tapped box was correct (only if Action Type is "Tap"), and the timestamp of the sample.

Line	Action	Position X	Position Y	Position Z	Touch X	Touch Y	Time Stamp	Cube ID	Correct?
...									
739	DRAG	325.7233	348.9596	205.2525	600	108	00:00:22.8033820		
740	DRAG	325.2326	348.8935	205.369	600	107	00:00:22.8196285		
741	DRAG	325.2326	349.3935	205.369	599	107	00:00:22.8699777		
742	DRAG	327.6911	349.1943	204.8247	779	328	00:00:22.1362298		
743	TAP	327.6912	349.1943	204.8247	779	328	00:00:23.1549178	Cube (15)	False
744	TAP	327.6912	349.1943	204.8247	788	366	00:00:23.8225771	Cube (7)	True
745	DRAG	322.2953	344.9468	206.1261	713	682	00:00:24.7858919		
746	DRAG	311.2677	335.8484	210.6286	725	659	00:00:24.8056675		
747	DRAG	305.8622	329.5708	214.8152	731	647	00:00:24.8196023		
...									

Table 5: Summary of accuracy statistics for both tests A and B of the 3D-Trail-Making Test. TP/FP Rates, Precision, Recall, and F-Measure shown are the weighted averages between the two classification labels.

	Accuracy	TP Rate	FP Rate	Precision	Recall	F-Measure
Test A	50 of 52 (96%)	0.96	0.04	0.96	0.96	0.96
Test B	45 of 51 (88%)	0.88	0.11	0.89	0.88	0.88

between X, Y, or Z direction of movement. For test A, this usage metric became important features in classifying the nationality of these participants. For test B, time played a slightly larger role, with the first and last time taken to tap on the correct box being an important feature while the total time taken in dragging in the X and Z directions being significant features also at the beginning and end of the examination.

Behavioral Observations. Observations on the participants while they were completing these tasks have yielded some interesting insights, which are supported by the most significant features chosen for classification. The Japanese participants visibly focused on completing the test as fast as possible. Although this did result in the participants moving the test sphere more rapidly, this did not result in faster time to completion, which is evidenced by the fact that total test completion time was not a significant feature for classification. However, this did result in a difference in behavior, which is evident in the classification features since velocity and the magnitude for the velocity vector were among the most significant features for classification. This may be attributed to a cultural difference due to the higher proliferation of mobile devices in Japan, and much higher rates of gaming mobile devices among the population [25]. Japanese participants tended to interpret this examination as a game, with the implicit objective being the fastest completion of time. This subtle cultural behavior was evident in the most important features, which in turn resulted in a feature set that yielded high accuracy and F-measure.

Test Accuracy Differences. A point of consideration are differences in classification features and accuracy between the tests A and B. We had already anticipated that the features for both tests would be considerably different, since the correct box order across both tests are entirely different like on the paper-based TMT. The considerable decrease in accuracy for test variation B can be explained by its increased difficulty, since participants across all forms of this test, paper or otherwise, frequently take a few moments between each number or letter to think which is the next correct box or dot in the sequence. We can observe this as most participants think aloud, repeating numbers and letters in sequence to themselves to remember. On this 3D-TMT version of test B we noticed that as they were thinking they would frequently move the sphere, often times in random directions as they would not yet know what box they were looking for. This has a cumulative effect of a considerably larger number of "random" movements in unpredictable directions in total when considering all 25 boxes, making participants slightly more difficult to classify based on movement patterns.

Establishing Behavioral Recognition. Overall, the classification accuracy across both tests A and B is high, strongly supporting the notion that this new 3D-TMT test, much like its paper-based counterpart, is sensitive to behavioral patterns. In this case, we can observe that the 3D-TMT is noticeably sensitive to a participant's geographic location and nationality. This classification task would be impossible with the traditional paper-based TMT, since a completed, fully connected set of dots would yield no insight as to the country of origin of its participant, and the paper's existing metric of noting completion time would also provide limited information into country of origin.

Limitations. The two quantitative studies and their respective results also highlighted areas of improvement. Although both the paper-based TMT and the novel 3D-TMT appear reasonably sensitive to geographic location and nationality, utility in broad behavioral recognition should firmly be established by conducting additional tests with patients with cognitive impairments. Additionally, we recognize the need for expanding the evaluation to determine whether the type of data gathered for this experiment may extend to general classification that also includes cognitive impairment.

5 Future Work

We are interested in continuing to explore the sensitivity of the 3D-TMT, testing for behavioral patterns such as those mentioned in Section 1.3. We seek to determine whether the test is sensitive in similar ways to the traditional TMT, such as age, level of education, sleepiness, inebriation, concussions, and Mild Cognitive Impairment among others. We also plan on integrating touch data into the classification features, including finger movement speed, location, acceleration, curvature, and motions indicating among others. This will introduce a new dimension of collected data which should even further increase the rich volumes of behavioral data that this examination can capture.

6 Conclusion

The 3-dimensional Trail-Making test is an examination heavily based on the principles of the established paper-based Trail-Making test with an added focus on collecting large volumes of behavioral data. We observe the participant's manipulation of the sphere object as a function of time with a high enough sample rate as to produce upwards of 2,000 data samples per test, which usually lasts less than one minute. The result is an ample set of features, 440 in total, and a rich set of data sensitive to subtle changes in behavior. To demonstrate this, we were able to classify completed tests as belonging to

either Japanese or American participants with an accuracy of up to 96.154%. We have also begun building our own normative data set for this new examination, with an improved skewness than that of established paper-based examinations. Future integration of touch data into the classification algorithm can provide an even higher volume of data likely to be more sensitive to behavioral patterns.

REFERENCES

[1] J. T. Barth, W. M. Alves, T. V. Ryan, S. N. Macciocchi, R. W. Rimel, J. A. Jane, and W. E. Nelson. Mild head injury in sports: Neuropsychological sequelae and recovery of function. *Mild head injury*, pp. 257–275, 1989.

[2] R. Canham, S. Smith, and A. Tyrrell. Location of structural sections from within a highly distorted complex line drawing. *IEE Proceedings-Vision, Image and Signal Processing*, 152(6):741–749, 2005.

[3] R. Canham, S. L. Smith, and A. M. Tyrrell. Automated scoring of a neuropsychological test: the rey osterrieth complex figure. In *Euromicro Conference, 2000. Proceedings of the 26th*, vol. 2, pp. 406–413. IEEE, 2000.

[4] S. Cavaco, A. Gonçalves, C. Pinto, E. Almeida, F. Gomes, I. Moreira, J. Fernandes, and A. Teixeira-Pinto. Trail making test: Regression-based norms for the portuguese population. *Archives of Clinical Neuropsychology*, 28(2):189–198, 2013.

[5] K. Cheshire, H. Engleman, I. Deary, C. Shapiro, and N. J. Douglas. Factors impairing daytime performance in patients with sleep apnea/hypopnea syndrome. *Archives of internal medicine*, 152(3):538–541, 1992.

[6] J. Davidoff and E. K. Warrington. The bare bones of object recognition: Implications from a case of object recognition impairment. *Neuropsychologia*, 37(3):279–292, 1999.

[7] B. Drew and J. Waters. Video games: Utilization of a novel strategy to improve perceptual motor skills and cognitive functioning in the non-institutionalized elderly. *Cognitive Rehabilitation*, 1986.

[8] P. Drotár, J. Mekyska, I. Rektorová, L. Masarová, Z. Smékal, and M. Faundez-Zanuy. Evaluation of handwriting kinematics and pressure for differential diagnosis of parkinson's disease. *Artificial intelligence in Medicine*, 67:39–46, 2016.

[9] R. J. Echemendia, M. Putukian, R. S. Mackin, L. Julian, and N. Shoss. Neuropsychological test performance prior to and following sports-related mild traumatic brain injury. *Clinical Journal of Sport Medicine*, 11(1):23–31, 2001.

[10] E. Frank, M. A. Hall, and I. H. Witten. *The WEKA Workbench. Online Appendix for Data Mining: Practical Machine Learning Tools and Techniques*. Morgan Kaufmann, fourth ed., 2016.

[11] M. Goebel. Ergonomic design of computerized devices for elderly persons-the challenge of matching antagonistic requirements. *Universal Acess in Human Computer Interaction. Coping with Diversity*, pp. 894–903, 2007.

[12] R. Z. Goldstein, A. C. Leskovjan, A. L. Hoff, R. Hitzemann, F. Bashan, S. S. Khalsa, G.-J. Wang, J. S. Fowler, and N. D. Volkow. Severity of neuropsychological impairment in cocaine and alcohol addiction: association with metabolism in the prefrontal cortex. *Neuropsychologia*, 42(11):1447–1458, 2004.

[13] F. Gong, W. Xu, J.-Y. Lee, L. He, and M. Sarrafzadeh. Neuroglasses: A neural sensing healthcare system for 3-d vision technology. *IEEE Transactions on Information Technology in Biomedicine*, 16(2):198–204, 2012.

[14] A. C. Hamdan and E. M. L. Hamdan. Effects of age and education level on the trail making test in a healthy brazilian sample. *Psychology & Neuroscience*, 2(2):199–203, 2009.

[15] R. Hashimoto, K. Meguro, E. Lee, M. Kasai, H. Ishii, and S. Yamaguchi. Effect of age and education on the trail making test and determination of normative data for japanese elderly people: the tajiri project. *Psychiatry and Clinical Neurosciences*, 60(4):422–428, 2006.

[16] C.-F. Jiang and Y.-S. Li. Development and verification of a vr platform to evaluate wayfinding abilities. In *Engineering in Medicine and Biology Society, 2009. EMBC 2009. Annual International Conference of the IEEE*, pp. 2396–2399. IEEE, 2009.

[17] H. B. Jimison, M. Pavel, K. Wild, P. Bissell, J. McKanna, D. Blaker, and D. Williams. A neural informatics approach to cognitive assessment and monitoring. In *Neural Engineering, 2007. CNE'07. 3rd International IEEE/EMBS Conference on*, pp. 696–699. IEEE, 2007.

[18] S.-W. Kim, J.-M. Kim, R. Stewart, K.-L. Bae, S.-J. Yang, I.-S. Shin, H.-Y. Shin, and J.-S. Yoon. Correlates of caregiver burden for korean elders according to cognitive and functional status. *International journal of geriatric psychiatry*, 21(9):853–861, 2006.

[19] M. Koso and S. Hansen. Executive function and memory in post-traumatic stress disorder: a study of bosnian war veterans. *European Psychiatry*, 21(3):167–173, 2006.

[20] R. Lara-Garduno, N. Leslie, and T. Hammond. Smartstrokes: digitizing paper-based neuropsychological tests. In *Revolutionizing Education with Digital Ink*, pp. 163–175. Springer, 2016.

[21] M. F. Mendez, R. L. Tomsak, and B. Remler. Disorders of the visual system in alzheimer's disease. *Journal of Neuro-Ophthalmology*, 10(1):62–69, 1990.

[22] M. Moetesum, I. Siddiqi, U. Masroor, and C. Djeddi. Automated scoring of bender gestalt test using image analysis techniques. In *Document Analysis and Recognition (ICDAR), 2015 13th International Conference on*, pp. 666–670. IEEE, 2015.

[23] M. Moetesum, I. Siddiqi, U. Masroor, N. Vincent, and F. Cloppet. Segmentation and classification of offline hand drawn images for the bgt neuropsychological screening test. In *Eighth International Conference on Digital Image Processing (ICDIP 2016)*, vol. 10033, p. 100334N. International Society for Optics and Photonics, 2016.

[24] H. B. Nazar, M. Moetesum, S. Ehsan, I. Siddiqi, K. Khurshid, N. Vincent, and K. D. McDonald-Maier. Classification of graphomotor impressions using convolutional neural networks: An application to automated neuro-psychological screening tests. In *Document Analysis and Recognition (ICDAR), 2017 14th IAPR International Conference on*, vol. 1, pp. 432–437. IEEE, 2017.

[25] S. Okazaki, R. Skapa, and I. Grande. Global youth and mobile games: applying the extended technology acceptance model in the usa, japan, spain, and the czech republic. In *Cross-Cultural Buyer Behavior*, pp. 253–270. Emerald Group Publishing Limited, 2007.

[26] D. Prvulovic, D. Hubl, A. Sack, L. Melillo, K. Maurer, L. Frölich, H. Lanfermann, F. Zanella, R. Goebel, D. Linden, et al. Functional imaging of visuospatial processing in alzheimer's disease. *Neuroimage*, 17(3):1403–1414, 2002.

[27] E. H. Seo, D. Y. Lee, K. W. Kim, J. H. Lee, J. H. Jhoo, J. C. Youn, I. Choo, J. Ha, and J. I. Woo. A normative study of the trail making test in korean elders. *International journal of geriatric psychiatry*, 21(9):844–852, 2006.

[28] W. Souillard-Mandar, R. Davis, C. Rudin, R. Au, D. J. Libon, R. Swenson, C. C. Price, M. Lamar, and D. L. Penney. Learning classification models of cognitive conditions from subtle behaviors in the digital clock drawing test. *Machine learning*, 102(3):393–441, 2016.

[29] E. Strauss, E. M. S. Sherman, and O. Spreen. *A Compendium of Neuropsychological Tests: Administration, Norms, and Commentary*. Oxford University Press, third ed., 2006.

[30] J. M. Thornbury. The use of piaget's theory in alzheimer's disease. *American Journal of Alzheimer's Care and Related Disorders & Research*, 8(4):16–21, 1993.

[31] T. N. Tombaugh. Trail making test a and b: Normative data stratified by age and education. *Archives of Clinical Neuropsychology*, 19(2):203–214, 2004.

[32] O. H. Turnbull, N. Beschin, and S. Della Sala. Agnosia for object orientation: Implications for theories of object recognition. *Neuropsychologia*, 35(2):153–163, 1997.

[33] O. H. Turnbull, D. P. Carey, and R. A. McCARTHY. The neuropsychology of object constancy. *Journal of the International Neuropsychological Society*, 3(3):288–298, 1997.

[34] I. Zavala-Ibarra and F. Jesús. Ambient videogames for health monitoring in older adults. In *Intelligent Environments (IE), 2012 8th International Conference on*, pp. 27–33. IEEE, 2012.

[35] P. Zham, D. K. Kumar, P. Dabnichki, S. P. Arjunan, and S. Raghav. Distinguishing different stages of parkinsons disease using composite index of speed and pen-pressure of sketching a spiral. *Frontiers in Neurology*, 8, 2017.