An Analysis of a Digital Variant of the Trail Making Test Using Machine Learning Techniques

Jessamyn Dahmen[a], Diane Cook[a], Robert Fellows[b], and Maureen Schmitter-Edgecombe[b]

[a]School of Electrical Engineering and Computer Sciences, Washington State University, Pullman, WA, USA

[b]Department of Psychology, Washington State University, Pullman, WA, USA

{jb3dahmen,djcook,robert.fellows,schmitter-e}@wsu.edu

Abstract

BACKGROUND:

The goal of this work is to develop a digital version of a standard cognitive assessment, the Trail Making Test (TMT), and assess its utility.

OBJECTIVE:

This paper introduces a novel digital version of the TMT and introduces a machine learning based approach to assess its capabilities.

METHODS:

Using digital Trail Making Test (dTMT) data collected from (N=54) older adult participants as feature sets, we use machine learning techniques to analyze the utility of the dTMT and evaluate the insights provided by the digital features.

RESULTS:

Predicted TMT scores correlate well with clinical digital test scores (r=0.98) and paper time to completion scores (r=0.65). Predicted TICS exhibited a small correlation with clinically-derived TICS scores (r=0.12 Part A, r=0.10 Part B). Predicted FAB scores exhibited a small correlation with clinically-derived FAB scores (r=0.13 Part A, r=0.29 for Part B). Digitally-derived features were also used to predict diagnosis (AUC of 0.65).

CONCLUSION:

Our findings indicate that the dTMT is capable of measuring the same aspects of cognition as the paper-based TMT. Furthermore, the dTMT's additional data may be able to help monitor other cognitive processes not captured by the paper-based TMT alone.

Keywords

Computerized cognitive assessment, design and validation, machine learning, mobile application, trail making test.

1. Introduction

Techniques to monitor cognitive health are essential to informing diagnostic decisions and effectively applying interventions early in the disease process [1]. Chronic neurodegenerative diseases that are characterized by cognitive decline, such as Alzheimer's and Parkinson's disease, increase in prevalence with age. These disease processes limit functional ability to perform activities of daily living important to overall health and well-being, such as cooking and bathing. Other aspects of an individual's life may also be affected. For example, as a result of memory loss or a decline in other cognitive abilities, social interaction may decrease leading to higher degrees of isolation in the older adult population. Currently, the resources and services that most care providers can offer may not be sufficient to address the needs of the coming "age wave" [2] and those impacted by cognitive decline. To confront this issue, researchers have utilized technology to help transform the way physical and mental health of the older adult population is monitored. The goal of this research is not to replace human care providers or clinicians. Rather, the goal is to augment the ability of care providers to effectively care for older adults in cost effective and non-intrusive ways.

In this study we introduce technology to monitor cognitive health.

Despite the wide variety of technologies that are already available, the most common method of measuring and monitoring cognitive health is paper-based cognitive assessment [3]. This traditional approach has several limitations [4]. Foremost, most traditional paper-based techniques are not designed for continuous repeated measurement or adaptability. The variety of data offered by paper-based techniques is limited and often fails to capture subtle, but crucial, periods of variability in performance. These assessments are often administered in lab environments for single time points, after an individual has already started to show signs of decline in cognitive ability. Due to their inability to adapt or offer new tasks to solve throughout multiple time points, many paper-based assessments may suffer from practice effects, where performance improves based on familiarity with the task. In addition, traditional methods can be expensive to administer and may not be practical with regard to accessibility for those in remote areas.

We introduce a digital, mobile variant of the original paper-based TMT, which we refer to as the digital TMT or dTMT. Rather than creating an entirely new cognitive test, the dTMT was designed to look as similar to the original test as possible in order to establish initial convergent validity with the paper-based version. Our dTMT was developed as an Android application and was administered using a capacitive touch screen tablet. Several sensor data based features were embedded into the design of the

dTMT without changing the layout or distracting the user. The dTMT is capable of outputting sensor data that may give deeper insights into certain aspects of cognition than paper-based tests. We build upon existing digital variants of the Trail Making Test [5, 6, 7, 8, 9] by including features such as detailed timing information, pauses, and lifts that are captured by utilizing a touchscreen interface. We then use these features in a machine learning based analysis. To our knowledge no other studies have been published that make use of the proposed digital TMT features in a machine learning based analysis. In this paper, we perform several experiments. First, we examine whether the dTMT can capture the same information as the paper-based TMT. Second, we assess whether machine learning techniques can be designed to automate the prediction of other neuropsychological standardized test scores based on the digital test. Third, we investigate whether the dTMT can provide additional insights not offered by pen-and-paper tasks that can be utilized to improve the performance of the machine learning-based automated assessment. Finally, we assess whether the dTMT data can be effectively applied to machine learning tasks with regard to classifying an individual as Healthy or Neurologic. The main contributions of this study are to introduce the design of a new digital variant of the TMT and to analyze the effectiveness and utility of the digital measures of performance using machine learning based techniques. We also assess the validity of the dTMT and explore how the

digital measures of performance may be used to provide clinicians with added information about cognitive performance when compared to a paper and pencil based approach. To validate the dTMT, we utilize machine learning methods to predict dTMT performance, paper TMT performance, and performance on other cognitive tests using various subsets of the digital TMT features. Our results indicate that the digital features do provide some correlation with these values and that the digital features overall improve this correlation. On the other hand, we observe that automatically classifying individuals as Healthy or Neurologic from dTMT features alone using a mixed neurologic participant group is a more challenging problem that will require additional research and incorporation of other digital performance-gathering tools.

2. Background

2.1. The Trail Making Test

The TMT is a standard neuropsychological test that has been extensively studied and used. It can assess a variety of neurological disorders [10]. The TMT is widely used in practice because it is freely available and can reliably identify several cognitive disorders. It has been shown to correlate with measures of speed, visuospatial skills, general fluid cognitive abilities, cognitive flexibility, set-switching, motor skills and dexterity [10,

11]. For these reasons we were motivated to use this test to develop a digital variant and build upon existing computerized versions of the TMT [5, 6, 7, 8, 9]. The paper-based version of the test has two parts, Part A and Part B. It is established that each condition of the test measures different cognitive processes. In Part A, the participant is instructed to draw a line connecting encircled numbers in ordered sequence (1,2,3. . . 26) as quickly as possible without lifting the pen from the surface of the paper. In Part B, the participant connects a series of circles containing either a number or letter in alternating sequence (1,A,2,B. . . 13). Each part is scored according to the total *time to completion* and *number of errors* committed. An error is recorded if the participant draws a line to a circle in the wrong order. Each participant is given 300 seconds to complete Part A and 300 seconds to complete Part B. If the participant did attempt to complete the TMT, but was unable to complete it in the given amount of time, the *time to completion* score is recorded as 301 seconds. In the paper-based version of the test each part is printed on a white piece of paper with black ink defining the circles, numbers, and letters. To draw lines connecting the circles the participant uses a pen. The scoring of the paper-based TMT, which includes overall time and number of errors, is recorded by the clinician administering the test.

2. 2. The TICS

The Telephone Interview for Cognitive Status (TICS) was originally designed to eliminate the need for face to face interviews. It is used as an overall brief screening assessment of cognitive function and does not require any motor or reading actions since it was designed to be administered over a telephone [12, 13]. Participants in our study completed all portions of the TICS over the phone as part of a screening process. The overall performance measure for the TICS is expressed as a single numerical score. This score can range from 0 - 41, where 25 and below represents impairment. This test is a useful screening measure of overall cognitive health and is insensitive to motor or reading abilities. In our experiments we use machine learning methods and our digital trail making task features to predict a TICS total performance score for each participant. We then compare the predicted scores with the TICS performance score obtained under clinical administration. This test was chosen to demonstrate our machine learning method's ability to predict overall cognitive health from the dTMT and the expanded set of captured features.

2.3. The FAB

The Frontal Assessment Battery (FAB) is a brief assessment tool used to assess executive functioning [14]. Participants in our study completed all

portions of the FAB in our lab environment. The overall performance measure for the FAB is expressed as a single numerical score. The total score ranges from 0 - 18, where a higher score indicates better performance. This test is especially useful for assessing significant executive dysfunction. In our experiments we use machine learning methods and our dTMT features to predict a FAB performance score for each participant. We then compare the predicted scores with the FAB performance score obtained under clinical administration.

2.4. Related Work on Digital Cognitive Assessment

Some paper-based methods can suffer from human error and bias, despite being standardized to maximize administration consistency. To address these limitations, researchers have applied technology to improve traditional paper-based approaches. These improved assessments are often called digital cognitive assessments or computerized cognitive assessments. In this paper we will refer to this technology as digital cognitive assessments or digital assessments. Digital assessments often incorporate constructs from existing paper-based assessments, some can be self-administered, and many have the capability to acquire additional data by utilizing interfaces on popular digital devices such as smart phones, computers, and tablets [4, 15, 16]. While some digital assessments require special equipment [17], many have been designed to be used with widely adopted commercial technologies

that older adults may already use to increase acceptance and usage. In addition, digital assessments have the potential to be designed to be adaptive and dynamic, thereby reducing practice effects [18]. For example, a digital assessment could be designed in the form of a series of game-like tasks that offers different levels and can take different skills and ability levels into consideration [19, 20, 21, 22, 23]. These advantages may enable digital assessments to be used as a tool to measure cognitive state over long periods of time in more natural environments. The benefits of digital assessments make them an ideal tool to help clinicians and care providers better assist older adults who may be at risk for cognitive decline. An example of a digital assessment is a computerized version of the Trail Making Test (TMT) developed by Salthouse and Fristoe. This digital assessment examined how looking at individual keystroke times in addition to total *time to completion* could help with performing a more analytical investigation of performance differences [7]. However, many existing digital assessment designs are often limited in terms of the performance measures they capture [16, 24].

3. The Design of the Digital Trail Making Test

The dTMT was developed using the Android operating system version 4.2.2 with the Android Development Tools Integrated Development Environment. The dTMT was designed to be an Android application running

on a 10.1 inch Samsung Galaxy Tab 2 capacitive touchscreen tablet. The dTMT was designed to be as consistent with the paper-based test as possible. Fig. 1 shows a screenshot of the dTMT Part A and Fig. 2 shows an image depicting the dTMT's interface.

The design of the dTMT was determined by iterative user tests with older adult participants aged 50 - 93 years, administered by a trained clinician. The dTMT underwent six design iterations based on feedback from participants and administrators, resulting in the final version we used for the data collection described in this paper. In addition to recording total *time to completion* in seconds and the *number of errors*, the dTMT also recorded several digital measures of performance that were used as features to train our machine learning models. These features were chosen in consultation with trained clinical neuropsychologists. They were selected based on their discriminatory power for clinical assessment as well as to effectively utilize information available from tablet sensors.

Time to Completion (Target Feature): The time in seconds it takes to draw a line connecting all circles in the correct order.

Number of Errors: The number of times a line is drawn to a circle in the incorrect order.

Average Pause Duration: The average duration in seconds of all the

pauses that occurred.

Average Lift Duration: The average duration in seconds of all the lifts of the stylus that occurred with regard to the tablet screen.

Average Rate Between Circles: Drawing rate between circles. Rate is defined as the straight line distance between the exit point of one circle and the entry point of the next, divided by the time spent drawing the line from one circle to the next.

Average Rate Inside Circles: The average drawing rate inside each circle.

Average Time Between Circles: Average time in seconds spent drawing between circles.

Average Time Inside Circles: Average time in seconds spent drawing inside circles.

Average Time Before Letters: The average drawing time in seconds before circles that contain letters (Part B only).

Average Time Before Numbers: The average drawing time in seconds before circles that contain numbers (Part B only).

Number of Pauses: Total number of pauses that occurred during the test.

Number of Lifts: Total number of lifts that occurred during the test.

Average Rate Before Letters: The average drawing rate before circles that contain letters (Part B only).

Average Rate Before Numbers: The average drawing rate before circles

that contain numbers (Part B only).

Total Average Pressure: Average pressure values are produced by the dTMT using Android hardware values. It is based on how hard the user is pressing on the screen using the surface area of the pressed stylus.

Total Average Size: Average size values are produced by the dTMT using Android hardware values. It is also based on how hard the user is pressing on the screen using the surface area of the pressed stylus.

In the paper version of the Trail Making Test when the user makes an error by drawing a line to a circle in the wrong order, the administrator will stop the participant and tell them to start from the last correct circle. The dTMT was designed with self-administration in mind. When the user makes an error the application will automatically draw a red X over the incorrect circle, visually informing the user that an error has been made. A visual example of the error indicator can be seen in Fig. 3a. However, in addition to informing the participant that the error has been made, the participant must also be redirected back to the last correct circle. For simplicity and consistency with the paper-based version of the test we then had the clinician direct the participant back to the last correct circle using the same method as in the paper-based test. In future designs of the dTMT it would be ideal to reorient the participant automatically using, for example, a combination of visual cues and audio instructions. However, a participant

with cognitive impairment might not respond to these automatic cues, so redirection by the clinician may be the most effective method for this test.

In the dTMT the participant uses a stylus to draw lines that connect circles. While the stylus was chosen to replicate a pen as much as possible the tip of the stylus is thicker than a pen tip. In many of our user trials the participants drew lines that looked extremely close to entering the circle but were not actually inside the circle based on the pixel precision of the application. This introduced some user frustration as the thicker stylus tip did not easily allow for precise drawing when compared to an actual pen. In order to compensate for this a tolerance of 5 pixels around the whole circumference of the circle was added. Fig. 3b demonstrates this tolerance area.

In order to record pauses and lifts as measures of performance several design decisions had to be made. In the case of pauses, a numerical threshold was determined with regard to what amount of time constituted a pause. In our experiments we determined a pause to be any time exceeding 0.1 seconds that the user held the stylus in the same place on the screen. The tolerance area of 5 pixels and this pause threshold of 0.1 seconds were determined based on clinical analysis of the dTMT during the initial six design iterations. In order to determine the stylus location on the screen we used the Android pointer events to programmatically assess if the user had touched the surface

of the screen with the stylus. These events are very sensitive to movement so even if the stylus appeared to be in the same place upon visual inspection the hardware would record natural trembling of the hand in healthy individuals as movement. To address this issue we also added a tolerance area to our movement detection method. From any touch location if the user did not move the stylus more than 5 pixels in any direction this was recorded as a pause. Fig. 3c demonstrates this approach. In addition to recoding pauses we also recorded when the user lifted up the stylus using the Android pointer events.

4. Methods

In our experiments all participants completed a battery of neuropsychological assessments, which included the original paper-based TMT, the dTMT, the FAB, and a number of other standard neuropsychological tests [25]. These assessments were administered by a trained clinician in a lab environment. The TICS was administered over the phone as part of an initial screening process. Two clinical neuropsychologists established diagnostic classification using results of these neuropsychological assessments, review of medical history, and clinical interview. Of note, the dTMT was not used in the diagnostic process.

During the testing session the paper TMT was always administered

first and approximately one hour separated the administration of the paper TMT and the dTMT. The tablet used to administer the dTMT was placed flat on the table in front of each participant. The participant was read the same set of instructions that are used in the paper version. Each participant was required to use the same pen-like stylus to complete the task. Similar to the paper based TMT, Part A and Part B of the dTMT started with an initial smaller practice version of Part A and Part B before the actual test was administered. After completing the Digital Trail Making Test, the captured data was then exported as CSV files onto an SD card inserted into the tablet. In these data files each participant is referenced by using a unique numerical identifier to ensure security and anonymity. The data was then stored on a secure server. A Python program was written to extract features from this raw data for use by the machine learning tools used in this study. Fig. 4 shows a high level view of the application design, flow, and data collection procedure.

Participants were 54 community dwelling older adults between the ages of 50 - 93. All participants provided informed consent. This was approved by the Washington State University Institutional Review Board (IRB) (Reference #12606-011). Older adult participants consisted of individuals who were cognitively healthy (N=28) and older adults diagnosed with neurological conditions including Mild Cognitive Impairment (MCI)

(N=6), Parkinson's disease (PD) (N=7), PD/MCI (N=3), and other (N=10). Table 1 shows the participant demographics in terms of diagnosis. Table 2 shows other participant demographics related to age, education, sex, and handedness. Diagnosis of MCI was consistent with criteria outlined by the National Institute on Aging Alzheimer's Association workgroup [26] and determined by performances on neuropsychological tests, self and informant interview data and medical records review. PD was diagnosed by a board certified neurologist in movement disorders and the other category were self reported conditions (e.g. traumatic brain injury, stroke) that may or may not have been confirmed by medical records and that resulted in cognitive difficulties.

In our experiments we use several standard regression based machine learning algorithms to predict scores. The algorithms used in this study are: Linear Regression (as a baseline), SMO support vector machine, and REPTree. We also used classification based machine learning algorithms to predict cognitive diagnoses with 10 fold cross validation. We used the WEKA implementation of these machine learning algorithms [27]. Each of these machine learning methods chosen vary in terms of their representation and underlying algorithms. There is no one best machine learning method to use so we utilize several widely used approaches and compare them.

To train and test our machine learning models we use a feature vector

comprised of the unique measures of performance described previously. Note that there are 16 individual features in our feature vector for dTMT Part B and 12 for dTMT Part A. Part A does not have letters so the time before letters and numbers features are excluded.

4.1. Validation

In this experiment we assess the validity of the dTMT. We also use this experiment to compare several regression based machine learning algorithms and then choose a regressor to use in later experiments based on performance. This method of validation was chosen to examine the validity of the dTMT from a different perspective than traditional correlation techniques and was based on the method utilized in 28].

We first use the dTMT measures of performance as features in each machine learning algorithm. We use these features to predict the paper-based TMT *time to completion* score. We then use the same approach to predict the dTMT *time to completion* score and compare these results to assess validity. We also use several feature subsets to further assess validity. By examining feature subsets we can identify how different features contribute to the *time to completion* score. This may provide clinicians with additional information about the cognitive factors that may impact TMT performance and influence the *time to completion* score.

The feature subsets used in this experiment are:

All Features: This set of features includes all the dTMT features described in the previous dTMT features section. This includes the digital *number of errors* but not the *time to completion* score.

Timing Features: This feature subset includes digital features related to timing. This includes times between and inside circles and rates between and inside circles.

Mobility Features: This feature subset includes digital features related to mobility. This includes information about pauses, lifts, pressure, and size.

4.2. Predicting the TICS and FAB scores

In this experiment we assess the utility of the alternative digital features. We explore how the features can be used to provide additional information beyond what the paper-based approach can provide. We use the dTMT and paper TMT measures of performance as features and predict the performance scores of two standard neuropsychological assessments, the TICS and FAB. We first use the dTMT features to predict the TICS and FAB performance scores using a regression based machine learning algorithm. We then compare the predicted scores with each participant's clinically derived score. We then use the same approach with only the paper-based

features from the paper-based test (*time to completion* and *number of errors*) to predict the TICS and FAB scores.

In this experiment we expand upon the digital feature subsets to include the *time completion* score for the dTMT. By examining these subsets we can also identify which features provide the most valuable information when predicting the scores of other neuropsychological tests. The expanded subsets include:

All Features: This set of features includes all the dTMT features described in the previous dTMT features section. This includes the digital *time to completion* and *number of errors*.

All Digital Features: The subset of features includes all the dTMT features except the digital *time to completion* and *number of errors*. This subset will help evaluate the value of the unique features captured by the digital task.

Standard Features: This feature subset includes only the digital *time to completion* and *number of errors*. This subset will help assess the value of the digital versions of features also captured by the paper TMT.

Timing Features: This feature subset includes digital features related to timing. This includes times between and inside circles, rates between and inside circles, and *time to completion*.

Mobility Features: This feature subset includes digital features related to mobility. This includes information about pauses, lifts, pressure, and size.

4.3. Classification of Healthy or Neurologic

In this experiment we explore how additional data captured by the dTMT can be applied to classification. We use the dTMT features and expanded digital feature subsets to classify each participant as Healthy (H) or Neurologic (N). To accomplish this, we use the Weka implementation of the SMO SVM, Naive Bayes, and J48 Decision Tree classifiers. We also evaluate the ability of these classifiers to accurately label each participant as CH or N based on 10-fold cross validation where each fold is randomly generated.

5. Results

5.1. Validation

Table 3 demonstrates the Pearson correlations between the predicted dTMT *time to completion* score and the clinically derived *time to completion* score using a Linear Regression, REP-Tree, and SMO-SVM algorithm respectively for all 54 older adult participants. Next, we repeat this experiment with the paper version of the test. The results are summarized in Table 4, which demonstrates the Pearson correlations between the predicted paper trails *time to completion* score and the clinically derived paper score for all 54 older adult participants.

5.2. Predicting the TICS and FAB scores

While the previous experiments focused on validating the machine learning algorithms by correlating predicted performance with actual performance, in this set of experiments we focus on correlating TMT performance with other cognitive tests. First, we learn a SMO-SVM model of TICS scores from subsets of dTMT features and evaluate the model by measuring correlation between the predicted and clinically derived TICS values, summarized in Table 5 for all 54 older adult participants.

In the previous experiment we determined that the SMO SVM machine learning algorithm demonstrated the best consistent performance. Next, we repeat these experiments using the FAB scores, as shown in Table 6. In order to compare learned models from the digital features and from paper-only features, we repeat the TICS experiment using only the paper trails score. Table 7 demonstrates the Pearson correlations between the predicted TICS total score and the clinically derived score using a SMO-SVM algorithm for all 54 older adult participants using only the paper trails *time to completion* and *number of errors* scores.

5.3. Classification of Healthy or Neurologic

Finally, we examine whether machine learning methods can utilize

dTMT features to automate classifying participants as Healthy or Neurologic. Table 8 shows both the percent accuracy (percent accuracy standard deviation is in parentheses) and area under the ROC curve value for classifying all 54 older adult participants as either Healthy or Neurologic using DTMT Part A feature subsets. Here we compute accuracy as the number of participants correctly classified divided by the total number of participants. Table 9 shows the results of the same experiment using dTMT Part B feature subsets.

6. Discussion

In our experiments we first considered whether machine learning models could predict TMT scores. As Tables 3 and 4 show, the models align with clinically-derived scores for both the digital and paper versions of the test. As expected, the predicted digital score in general results in higher correlation than the predicted paper score, which is intuitive given that the model is learned from digital features. Furthermore, the amount of correlation is not consistent across all learning algorithms. While SMO-SVM performs the best overall, especially in Table 3, the Linear Regression algorithm does outperform this algorithm for individual cases in Table 4. In Table 3 we see that in the case of SMO and Linear Regression the All features subset results in stronger correlation than the other subsets for Part

A, but this is not the case for Linear Regression for Part B. In Table 4 we see more variability, where the All features subset does not always result in the strongest correlation for each algorithm. One might anticipate that using the All features subset should consistently perform the best, but this is not always the case when features are included that might actually turn out to be poor indicators of overall performance.

We next turn our attention to predicting scores on other cognitive tests. As Tables 5 and 6 show, no large correlations were found for any of the feature subsets. In Table 6 we see slightly stronger correlations than found in Table 5 for Part B. One explanation for this is that Part B of the TMT is thought to measure executive function which the FAB has been shown to measure. We found that in general the dTMT features did result in stronger correlations than using the paper TMT features alone as shown in Table 7. This suggests that the dTMT features may add some amount of more useful information for machine learning applications than the paper-based features alone. The weak correlations overall for both predicting the TICS and FAB performance scores suggest that the current available features are not sufficient for effectively predicting the performance scores of these other cognitive tests. In future studies, tests measuring cognitive constructs closer to those being captured by the dTMT (e.g., set-switching, processing speed) could be examined. An additional important area of research is to examine how well

the dTMT features can predict everyday functioning.

Finally, we use machine learning methods to classify participants as Healthy or Neurologic from the dTMT features. The Area Under the ROC Curve-based results, shown in Tables 8 and 9, indicate that the classification is difficult to learn from dTMT features alone. However, for some feature subsets and classifiers the performance is better than what would be obtained by randomly guessing the values which would result in a AUC of around 0.5. In Table 8 we see that Naive Bayes performs better with subsets of features that include the majority of available dTMT features (All and All Digital) or features related to timing. In the case of SMO the Timing features subset results in the highest performance. In Table 9 we observe that using the Mobility features alone in general, results in lower performance than using the Timing and All features subset for all classifiers. These results suggest that Healthy or Neurologic can be learned to an extent based on the dTMT. They also indicate that features related to timing are slightly more informative when applied to this learning task than other features. The participants in our experiments who were diagnosed with neurological conditions are very heterogeneous with respect to their exact neurological diagnosis. Furthermore, the number of cognitively healthy participants and the number of participants with neurological conditions is not well-balanced. These values might improve with a larger, more balanced, more homogeneous sample and

would very likely improve if we included digital features from other cognitive tests as well.

7. Conclusions

In this paper we introduce a digital version of a traditional cognitive test, the Trail Making Test. The digital version of the test facilitates easier and more wide-spread test administration. In addition, the digital media allows us to capture a greater set of performance features than can be easily observed using the paper test.

To examine the dTMT, we utilize machine learning methods to predict dTMT performance, paper TMT performance, and performance on other cognitive tests using various subsets of the digital TMT features. Our results indicate that the digital features do provide some correlation with these values and that the digital features overall improve this correlation. On the other hand, we observe that automatically classifying individuals as Healthy or Neurologic from dTMT features alone using a mixed neurologic participant group is a more challenging problem that will require additional research and incorporation of other digital performance-gathering tools.

8. Acknowledgment

The authors would like to thank the National Science Foundation for supporting this research under grant number DGE-0900781.

9. References

[1] Tornatore JB, Hill E, Laboff JA, McGann ME. Self-administered screening for mild cognitive impairment: initial validation of a computerized test battery. The Journal of neuropsychiatry and clinical neurosciences. 2005;17(1):98–105.

[2] Picchi A. So many elderly, so few care workers. CBS Money Watch. 2015 July 14.

[3] Rabin LA, Paolillo E, Barr WB. Stability in test-usage practices of clinical neuropsychologists in the United States and Canada over a 10-year period: A follow-up survey of INS and NAN members. Archives of Clinical Neuropsychology. 2016 Mar 16;31(3):206-230.

[4] Bauer RM, Iverson GL, Cernich AN, Binder LM, Ruff RM, Naugle RI. Computerized neuropsychological assessment devices: joint position paper of the American Academy of Clinical Neuropsychology and the National Academy of Neuropsychology. The Clinical Neuropsychologist. 2012;26(2):177–196.

[5] Makizako H, Shimada H, Park H, Yoshida D, Uemura K, Tsutsumimoto K, et al. Evaluation of multidimensional neurocognitive function using a tablet personal computer: Test–

retest reliability and validity in community-dwelling older adults. Geriatrics & gerontology international. 2013;13(4):860–866.

[6] Onoda K, Hamano T, Nabika Y, Aoyama A, Takayoshi H, Nakagawa T, et al. Validation of a new mass screening tool for cognitive impairment: Cognitive Assessment for Dementia, iPad version. Clinical interventions in aging. 2013;8:353.

[7] Salthouse TA, Fristoe NM. Process analysis of adult age effects on a computer-administered Trail Making Test. Neuropsychology. 1995;9(4):518.

[8] Zakzanis KK, Mraz R, Graham SJ. An fMRI study of the trail making test. Neuropsychologia. 2005;43(13):1878–1886.

[9] Woods DL, Wyma JM, Herron TJ, Yund EW. The effects of aging, malingering, and traumatic brain injury on computerized trail-making test performance. PloS one. 2015 Jun 10;10(6):e0124345.

[10] Salthouse TA. What cognitive abilities are involved in trail-making performance?. Intelligence. 2011 Aug 31;39(4):222-32.

[11] Strauss E, Sherman EM, Spreen O. A compendium of neuropsychological tests: Administration, norms, and commentary. Oxford University Press, USA; 2006.

[12] Moylan T, Das K, Gibb A, Hill A, Kane A, Lee C, et al. Assessment of cognitive function in older hospital inpatients: is the Telephone Interview for Cognitive Status (TICS-M) a useful alternative to the Mini Mental State Examination? International journal of geriatric psychiatry. 2004;19(10):1008–1009.

[13] de Jager CA, Budge MM, Clarke R. Utility of TICS-M for the assessment of cognitive function in older adults. International journal of geriatric psychiatry. 2003;18(4):318–324.

[14] Lima CF, Meireles LP, Fonseca R, Castro SL, Garrett C. The Frontal Assessment Battery (FAB) in Parkinsons disease and correlations with formal measures of executive functioning. Journal of neurology. 2008;255(11):1756–1761.

[15] Tong T, Chignell M. Developing a Serious Game for Cognitive Assessment: Choosing Settings and Measuring Performance. In: Proceedings of the Second International Symposium of Chinese CHI. Chinese CHI '14. New York, NY, USA: ACM; 2014. p. 70–79. Available from: http://doi.acm.org/10.1145/2592235.2592246.

[16] Brouillette RM, Foil H, Fontenot S, Correro A, Allen R, Martin CK, et al. Feasibility, reliability, and validity of a smartphone based application for the assessment of cognitive function in the elderly. PloS

one. 2013;8(6):e65925.

[17] Davis R, Libon DJ, Au R, Pitman D, Penney DL. THink: Inferring Cognitive Status from Subtle Behaviors. InAAAI 2014 Jun 21 (pp. 2898-2905).

[18] Wouters H, Zwinderman AH, van Gool WA, Schmand B, Lindeboom R. Adaptive cognitive testing in dementia. International journal of methods in psychiatric research. 2009;18(2):118–127.

[19] Amato C, Shani G. High-level reinforcement learning in strategy games. In: Proceedings of the 9th International Conference on Autonomous Agents and Multiagent Systems: volume 1-Volume 1. International Foundation for Autonomous Agents and Multiagent Systems; 2010. p. 75–82.

[20] Cowley B, Kosunen I, Lankoski P, Kivikangas JM, Järvelä S, Ekman I, et al. Experience assessment and design in the analysis of gameplay. Simulation & Gaming. 2013;45(1):41-69.

[21] Liu C, Agrawal P, Sarkar N, Chen S. Dynamic difficulty adjustment in computer games through real-time anxiety-based affective feedback. International Journal of Human-Computer Interaction. 2009;25(6):506–529.

[22] Andrade G, Ramalho G, Santana H, Corruble V. Automatic computer game balancing: a reinforcement learning approach. In:

Proceedings of the fourth international joint conference on Autonomous agents and multiagent systems. ACM; 2005. p. 1111–1112.

[23] Jimison HB, Pavel M, Bissell P, McKanna J. A framework for cognitive monitoring using computer game interactions. InMedinfo 2007: Proceedings of the 12th World Congress on Health (Medical) Informatics; Building Sustainable Health Systems 2007 (p. 1073). IOS Press.

[24] Gualtieri CT. Dementia screening using computerized tests. Journal of Insurance Medicine. 2004;36:213–227.

[25] Schmitter–Edgecombe M, McAlister C, Weakley A. Naturalistic assessment of everyday functioning in individuals with mild cognitive impairment: The day-out task. Neuropsychology. 2012 Sep;26(5):631.

[26] Albert MS, DeKosky ST, Dickson D, Dubois B, Feldman HH, Fox NC, et al. The diagnosis of mild cognitive impairment due to Alzheimer's disease: Recommendations from the National Institute on Aging-Alzheimer's Association workgroups on diagnostic guidelines for Alzheimer's disease. Alzheimer's & Dementia. 2011;7(3):270–279.

[27] Hall M, Frank E, Holmes G, Pfahringer B, Reutemann P, Witten IH. The WEKA data mining software: an update. ACM SIGKDD explorations newsletter. 2009;11(1):10–18.

[28] Dawadi PN, Cook DJ, Schmitter-Edgecombe M. Automated cognitive health assessment using smart home monitoring of complex tasks. Systems, Man, and Cybernetics: Systems, IEEE Transactions on. 2013 Nov;43(6):1302-13.

Table 1: Participant Diagnosis Distribution

Diagnosis	Number	%
MCI	6	11%
Healthy	28	52%
PD/MCI	3	5%
PD	7	13%
Other	10	19%

Table 2: Participant Demographics

Demographic	Mean (SD)
Age	68.4 (9.3)
Education	16.2 (2.5)
Sex (% Female)	80
Handedness (% Right)	96

Table 3: Pearson correlation between predicted dTMT *time to completion* score and clinically derived dTMT *time to completion* score using various regressors for all older adult participants[1].

Feature Subsets	Linear Regression	REP-Tree	SMO-SVM
All (Part A)	0.97*ǂ	0.70*ǂ	0.98*ǂ
Timing (Part A)	0.83*ǂ	0.33*ǂ	0.88*ǂ
Mobility (Part A)	0.93*ǂ	0.71*ǂ	0.93*ǂ
All (Part B)	0.82*ǂ	0.79*ǂ	0.96*ǂ
Timing (Part B)	0.89*ǂ	0.67*ǂ	0.93*ǂ
Mobility (Part B)	0.95*ǂ	0.69*ǂ	0.95*ǂ

* indicates the correlation was statistically significant with $p < 0.05$. ǂ indicates the correlation was statistically significant with $p < 0.05$ with Bonferroni correction for n tests.

Table 4: Pearson correlation between predicted paper trails *time to completion* score and clinically derived paper trails *time to completion* score using various regressors for all older adult participants[1].

Feature Subsets	Linear Regression	REP-Tree	SMO-SVM
All (Part A)	0.60*ǂ	0.32*ǂ	0.65*ǂ
Timing (Part A)	0.60*ǂ	0.01	0.66*ǂ
Mobility (Part A)	0.63*ǂ	0.25	0.64*ǂ
All (Part B)	0.58*ǂ	0.27*ǂ	0.51*ǂ
Timing (Part B)	0.58*ǂ	0.34*ǂ	0.53*ǂ
Mobility (Part B)	0.50*ǂ	0.41*ǂ	0.54*ǂ

1. * indicates the correlation was statistically significant with $p < 0.05$. ǂ indicates the correlation was statistically significant with $p < 0.05$ with Bonferroni correction for n tests.

Table 5: Pearson correlation between predicted TICS total score and clinically derived TICS total score using an SMO-SVM for all older adult participants[1].

Feature Subsets	Part A	Part B
All	0.12	0.10
All Digital	0.11	0.01
Standard	0.11	0.33*
Timing	0.04	0.10
Mobility	0.02	0.23

1. * indicates the correlation was statistically significant with $p < 0.05$. No correlation was significant with $p < 0.05$ with Bonferroni correction for n tests.

Table 6: Pearson correlation between predicted FAB total score and clinically derived FAB total score using an SMO-SVM for all older adult participants[1].

Feature Subsets	Part A	Part B
All	0.13	0.29
All Digital	0.13	0.30*
Standard	0.10	0.23
Timing	0.23	0.41* ǂ
Mobility	0.04	0.01

1. * indicates the correlation was statistically significant with $p < 0.05$. ǂ indicates the correlation was statistically significant with $p < 0.05$ with Bonferroni correction for n tests.

Table 7: Pearson correlation between predicted TICS total score and clinically derived TICS total score using an SMO-SVM and only the paper trails *time to completion* and *number of errors* scores for all older adult participants[1].

	Part A	Part B
TICS	0.08	0.14
FAB	0.11	0.18

1. None of the correlations were significant ($p < 0.05$).

Table 8: Part A percent accuracy (Acc) and Area Under the ROC Curve (AUC) for all older adult participants using three different classifiers.

Feature Subsets	Naive Bayes		J48 Tree		SMO-SVM	
	Acc (SD)	AUC	Acc (SD)	AUC	Acc (SD)	AUC
All	62.6 (19.2) %	0.68	49.0 (19.0)%	0.47	53.3 (19.6)%	0.53
All Digital	65.1 (19.3)%	0.69	44.4 (17.5)%	0.42	52.4 (19.3)%	0.52
Standard	56.0 (19.1)%	0.67	52.1 (13.2)%	0.52	56.6 (13.9)%	0.55
Timing	67.9 (18.4)%	0.73	48.8 (14.8)%	0.49	62.8 (18.0)%	0.63
Mobility	59.2 (18.5)%	0.58	43.7 (13.8)%	0.41	55.3 (15.1)%	0.54

Table 9: Part B percent accuracy (Acc) and Area Under the ROC Curve (AUC) for all older adult participants using three different classifiers.

Feature Subsets	Naive Bayes		J48 Tree		SMO-SVM	
	Acc (SD)	AUC	Acc (SD)	AUC	Acc (SD)	AUC
All	59.4 (18.6) %	0.58	61.5 (16.6)%	0.63	57.7 (16.9)%	0.57
All Digital	60.7 (18.5)%	0.58	61.3 (16.4)%	0.63	56.1 (17.2)%	0.56
Standard	64.6 (17.1)%	0.64	58.1 (14.0)%	0.57	56.4 (12.2)%	0.55
Timing	64.7 (17.3)%	0.63	62.1 (16.4)%	0.63	59.3 (17.0)%	0.59
Mobility	57.6 (16.9)%	0.56	46.4 (11.3)%	0.42	56.5 (17.0)%	0.55

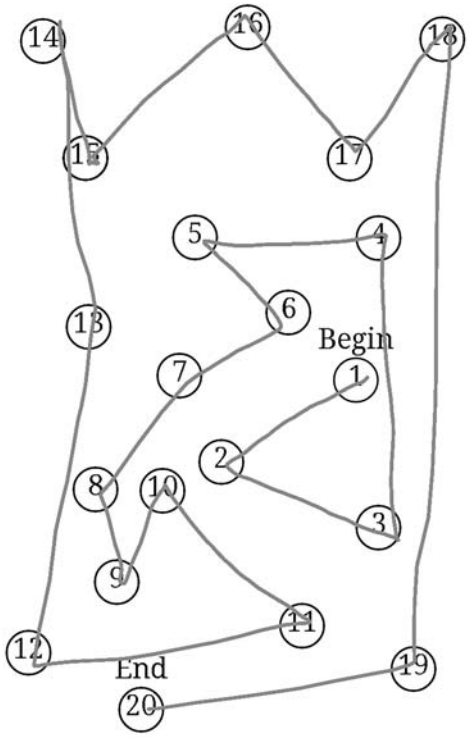
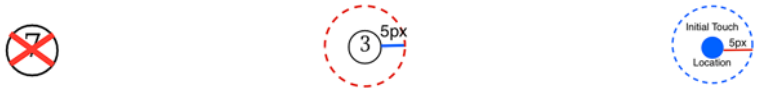


Figure 1: A visual example of the digital Trail Making Test Part A. Participants draw lines connecting the circles. Due to tablet size constraints 20 circles were used for Part A.

Figure 2: A visual demonstration of the Digital Trail Making Test's interface.



(a) Error (b) Tolerance (c) Pause Tolerance

Figure 3: dTMT design decisions visual demonstrations.

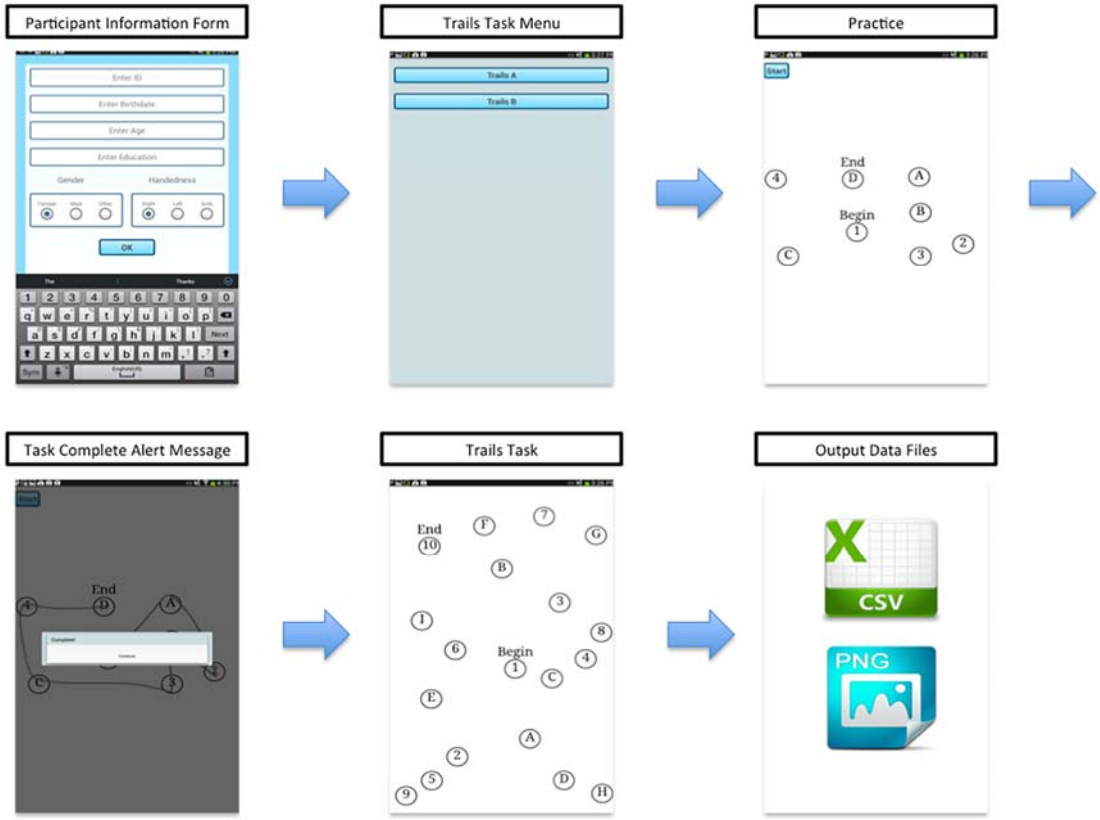


Figure 4: The high level design of the Digital Trail Making Test System.