Multi-digit Processing and Contextualized Analysis on the Digital Symbol Digit Test

by

Elizabeth A. DeTienne

Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

May 2020

Author . . . .
Department of Electrical Engineering and Computer Science
May 18, 2020

Certified by . . . .
Randall Davis
Professor
Thesis Supervisor

Certified by . . . .
Dr. Dana L. Penney
Director of Neuropsychology, Lahey Hospital & Medical Center
Thesis Supervisor

Accepted by . . . .
Katrina LaCurts
Chair, Master of Engineering Thesis Committee

Multi-digit Processing and Contextualized Analysis on the Digital Symbol Digit Test

by

Elizabeth A. DeTienne

Submitted to the Department of Electrical Engineering and Computer Science
on May 18, 2020, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Neurocognitive decline has been shown to occur as early as 15-20 years before obvious symptoms develop for patients with Alzheimer's. Current therapies work best when begun early, but it is very difficult to detect subtle cognitive change before obvious symptoms manifest. We have done analysis on the handwritten data from the digital Symbol Digit Test with the aim of detecting subtle cognitive decline early in the disease progression. We used a large dataset to augment the MNIST handwritten digit dataset. This has enabled us to recognize digits 0-12 with high accuracy, making it possible to automate scoring of the test. We also analyzed subtle features of the handwriting. We contextualized this data through visualizations, revealing a number of interesting trends and deviations for healthy patients versus patients with cognitive decline. For example, impaired participants tend to have more ink than we would expect for their average digit height, and pause for longer before writing a digit. We believe that this analysis will provide valuable new insights into a person's cognitive status.

Thesis Supervisor: Randall Davis
Title: Professor

Thesis Supervisor: Dr. Dana L. Penney
Title: Director of Neuropsychology, Lahey Hospital & Medical Center

Acknowledgments

Thank you to Professor Davis and Dr. Penney, for being such strong mentors to me. I appreciate that they have always been willing to offer guidance or wisdom, in research and beyond. In the midst of a pandemic, they were both supportive and caring in a variety of aspects of my life. They have each helped me grow tremendously as a researcher, engineer, and communicator. Thank you also to my family and friends, who have been there for me at every step along the way.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Neurodegenerative diseases, such as Alzheimer's Disease and Parkinson's Disease, are a major problem worldwide. Changes in the brain begin as early as 15-20 years before symptoms manifest in the patient. Current treatments are not able to cure these diseases, but are able to slow the progression or improve functioning. Treatments are more effective to improve quality of life the earlier they are begun. However, detecting the disease before obvious symptoms arise is a longstanding problem. We believe that our work with the digital Symbol Digit Test will help enable earlier recognition of cognitive decline.

Our work started with tackling the problem of recognizing handwritten two-digit numbers. Previous work relied on manual grading in order to score a dSDT [11]. We were able to construct a machine learning model to recognize the digits between 0 and 12 with 99% accuracy. The ability to automatically interpret written responses and assess the confidence for each prediction improves scoring accuracy and efficiency.

We extracted a variety of interesting features using the data from a digitizing pen. We looked at which boxes were revisited, where revisiting of a box means that a patient writes in some response box, then moves on, and then returns to the original box. Reasons for revisiting could include correcting mistakes. Data from the pen consists of timestamped coordinates on the page. We were able to use that data to determine which boxes are revisited.

We analyzed the amount of ink that the patient used. Increased ink within a response box may indicate crossing out a mistake, while increased total test ink may reflect large handwriting, or increased errors.

We were also able to create a list of boxes that were started out of order. For example, if a patient skips an answer box, it may indicate that they had trouble remembering the symbol-digit pairing. Conversely, it may indicate that they are strategically completing the test by answering similar questions together, even if they are not spatially next to each other (even though the instructions state to complete the test sequentially). In the delayed recall task, order indicates which pairings the subject remembered the most or least.

We detect and record how long the patient pauses before starting to write in each response box. Consistently long pauses across a test may indicate that the patient was experiencing a high cognitive load to complete the test, which may be an indicator of cognitive decline. We also computed a *think time* versus *ink time* metric. This describes how much of the time during the test was pen-to-paper (ink time) or in air (presumably time spent thinking), extending our work on similar variables that have proven sensitive to parsing motor and cognitive impairments in motor disorders such as Parkinson's Disease.

We computed response location and location drift or change over the course of a test. It's possible that unusual spatial deviations may be an indicator of cognitive decline, for example if the patient has trouble mentally shifting from one response box to the next.

Our next stage of research was contextualizing this data. We wanted to answer the questions, "What is normal? What is abnormal? What are good indicators of cognitive decline, or of specific neurocognitive diseases?" The analyses in Chapter 4 attempt to answer these questions.

Chapter 2

Background

2.1 The digital Symbol Digit Test

One of the ways that physicians test for cognitive decline is with a variety of handwritten tests that require patients to answer a series of questions. The digital Symbol Digit Test (dSDT), developed as a collaboration between Lahey Medical Center and MIT, is one test of this sort (Figure 2-1). It is a rapid graphomotor task comprised of three tasks: translation, copy, and delayed recall. The translation task of the test comes first. At the top of each task, there is a printed key that gives a one-to-one mapping of a symbol to a digit. The translation portion of this test is a series of data boxes, filled with symbols that correspond to the key. For each data box, there is an empty response box. As a person takes this test, they are expected to write in the corresponding response box whichever digit matches the symbol in the data box.

The copy task has a very similar format, but the patient is simply to copy the digit they see in the data box into the response box. This portion of the digital Symbol Digit Test gives a measure of the patient's performance under low cognitive load and serves as a structured delay before recalling the symbol-digit pairings.

The delayed recall task is last, with six symbols whose corresponding digits must be recalled from memory, as the symbol-digit key is not visible. This is a test for how well the patient may have learned the symbol-digit pairings, after the delay of completing the copy task.

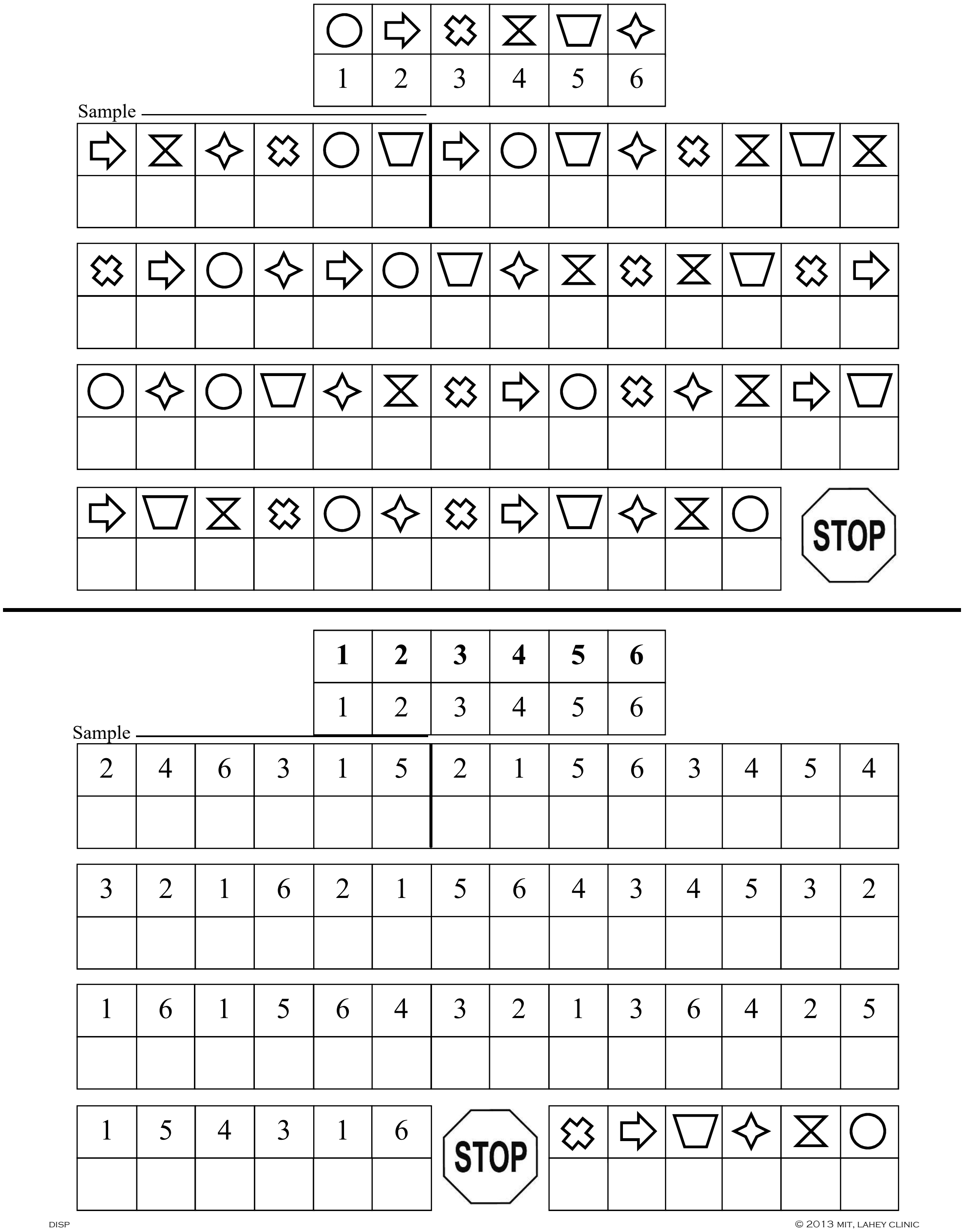

Figure 2-1: Display Version of the dSDT

Both the translation and copy tasks begin with a six item sample section that includes one example of each symbol-digit pair. This gives the test-taker an opportunity to become familiar with the task and have equal exposure to each item.

The written response data of a patient is collected using a digitizing pen developed by Anoto, Inc. This pen functions the same as any other writing utensil, but in addition records the pen strokes electronically. Inside the pen, there is a camera aligned with the ink barrel which records the timestamp and location coordinates of points within each pen stroke on the surface of the test form.

2.2 Medical Background

For many neurocognitive disorders change is insidious, but for Alzheimer's Disease, changes in the brain of a patient may occur 15-20 years before symptoms of cognitive decline become outwardly obvious [19]. Each stage of the disease progression is marked by the effect it has on the patient's independence, as well as their social and occupational skills. Symptoms will present in terms of memory loss, other cognitive function degradations, changes in demeanor or sleep, and more.

Current cognitive assessment tools detect errors when cognitive change is apparent but miss Alzheimer's Disease in the preclinical stage. Current therapies work best when they are begun early in the disease progression, and delayed detection can result in cognitive impairment that can be delayed or someday prevented. Our research is aimed at filling this need.

We extracted features from the digital Symbol Digit Test that are indicative of cognitive status. We evaluated the effectiveness of these features based on diagnosis information as well as scores on two well-known cognitive screening tests: the Montreal Cognitive Assessment (MoCA) and the Mini-Mental State Examination (MMSE). Current research has shown both the MoCA and MMSE are effective at detecting cognitive impairment ([14], [9]). Both are scored from 0-30, with 30 being the healthiest score, and scores below 26 indicating increased cognitive impairment [5]. The MMSE ([7], [10]) has two variations: the world version, and the serial sevens version. Both the MoCA and the MMSE are administered by a trained clinician. Research suggests that the MoCA may be superior to the MMSE for earlier detection of cognitive impairment [14]. In our dataset, we have some subjects with MoCA scores, and others with either of the variations of the MMSE.

The core of our project is detecting subtle cognitive changes through the administration of a handwritten test. We were guided in part by existing research on how various neurodegenerative diseases affect handwriting [24]. This paper gave us inspiration to look at things like centeredness of the response within the response box.

Work in [13] analyzes handwriting in shapes and words for subjects with Alzheimer's Disease or Parkinson's Disease. This research also used an electronic pen, and focused primarily on the acceleration and velocity of pen strokes. Unfortunately they had a limited dataset. In our research, we were able to extract a larger set of features from a larger dataset. Also instead of free-form writing data, our data is from a well-structured test.

2.3 Technical Background

Song's paper [21] proves a foundation for interpreting spatio-temporal data from a digitizing pen. This paper discusses unwinding the handwritten strokes, so that we can uncover what was written at each layer independently. This is an advantage of digitizing pen data – it enables the understanding of behavior that would not be apparent from overwritten strokes. While this work was insightful, it was tackling a slightly different problem than ours. Song's work used the Clock Drawing Test, and focused on unstructured spatial analysis such as distinguishing the clock hands, the clock digits, and the circle around the clock from each other. In the digital Symbol Digit Test, the user writes individual answers in individual boxes. This makes our task of isolating digits much easier.

Huang [11] generated a rich array of promising leads, such as centeredness and prebox pause. The current work advances our understanding of these features.

Several papers, projects, and tutorials were helpful in developing our digit recognizer ([3], [8], [22], [25]). Recognition of handwritten single-digit numbers is basically considered a solved problem, and many of these resources share models that report 99% accuracy or higher. Keras is a well-known and reliable resource in the machine learning community, so we decided to go directly to the source while creating our two-digit recognizer. We ended up using the digit recognizer model from the Keras Team itself [3] (which was built for single digit classification) to build a two-digit classifier.

In doing this we used the MNIST, a group of handwritten single-digit numbers with 60,000 training examples and 10,000 testing examples [6]. Algorithms using this dataset have consistently reported an accuracy above 99% across multiple approaches.

Chapter 3

Methods

3.1 Machine Learning to Classify Digits 0-12

The first part of this project was classifying the subjects' handwritten strokes from the test as indicating one of the digits between 0 and 12. We started with the high-performance models already trained on the MNIST dataset, but these models do not address two-digit numbers. We created a two-digit classifier that could meet our needs.

3.1.1 Ideas and Alternatives Considered

We considered segmenting the pen data stroke-by-stroke with the aim of isolating each digit, then interpreting each digit alone. However, ordinary handwriting can be hard to read, particularly when writing quickly. For example, elevens can look like backwards capital N's when people fail to lift their pen between digits. Digits can also appear different depending upon context: the leading 1 in a two-digit number may look different than a standalone 1. For these reasons, we decided not to use segmentation.

We considered using existing open-source code to generate training data for two-digit numbers. Several repositories on Github ([2], [20]) used MNIST to generate a series of two-digit numbers by juxtaposing two single-digit numbers. While this was

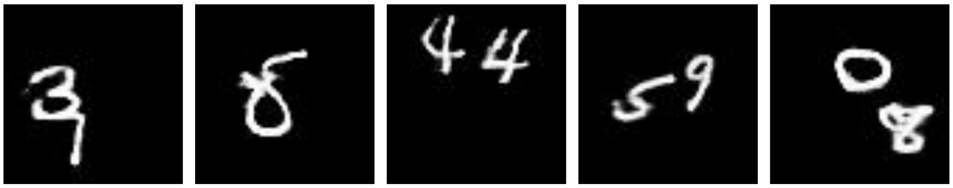

Figure 3-1: External Repository's Two-Digit Generator Results [2]

a plausible idea, it produced several problems. For example, the training set did not accurately represent how people write two-digit numbers. As evident in Figure 3-1, many of the created numbers are oddly vertically offset. Also as noted, people tend to write digits differently when they write multiple digits continuously. Hence, we can't simply compose single-digit numbers together to try to create double-digit numbers.

3.1.2 Our Solution

As a result, we decided to find handwritten instances of the digits 10, 11, and 12 to use along with the MNIST digits to train the model. The dSDT test has been used in a multi-center consortium, including the Lahey Clinic in Massachusetts and the University of Florida. Between all sites of administration of the test, we have nearly 900 completed trials of the dSDT. As each trial contains 18 examples of each digit, we have in principle 16,200 examples of each of the digits 0-12.

We examined in each trial all locations where a 10, 11, or 12 was expected, to find the handwritten instances that were correct and legible. We were able to create a superset of MNIST so that we had handwritten digits for 0-12.

Digital pen data is recorded as time-stamped locations, but we wanted our data to look similar to the original MNIST data. We used matplotlib to save each response as an image using `plt.plot` to draw them. This also automatically centers the digit, which is similar to the way digits were centered in the MNIST dataset. We set the size of each image to 28x28, while preserving the height-to-width ratio of how each digit was originally drawn. We plotted each penstroke in grayscale and with a linewidth to visually match MNIST images. We also normalized our pixel values to the range 0-1 to match the MNIST data format.

Table 3.1: Size of Dataset for Each Digit

Digit	Training Data Size	Test Data Size	Data Source
0	5923	980	MNIST
1	6742	1135	MNIST
2	5958	1032	MNIST
3	6131	1010	MNIST
4	5842	982	MNIST
5	5421	892	MNIST
6	5918	958	MNIST
7	6265	1028	MNIST
8	5851	974	MNIST
9	5949	1009	MNIST
10	5977	1023	digital Symbol Digit Test
11	6034	966	digital Symbol Digit Test
12	5989	1011	digital Symbol Digit Test
Totals	78,000	13,000	

We then trained the model, reusing the architecture of a model built for the original MNIST dataset [3] (see model architecture in Appendix). The resulting system had an accuracy above 99% for digits 0-12, which compares favorably with the best of the single-digit classifiers online.

3.2 Extracting Qualities of Interest

We extracted a series of features that might be useful in detecting early cognitive decline.

3.2.1 Digit prediction and confidence for each cell

For each dSDT response box, we used the neural net described above to make a prediction of what digit was written in it, along with a confidence estimate. A low confidence (like 0.40 or lower) means the neural net is not sure of any prediction. Empirically, it's likely that whatever was written is unclear or outside the set of digits 0-12. We expect that displaying the confidence will allow clinicians to weight interpretation accordingly in aiding diagnosis.

3.2.2 Order of boxes touched

We extracted the order of the boxes touched. We use *touched* to mean that the pen touched the paper within the bounds of a single response box. This metric may be useful to see if the patient jumped around, skipped ahead, went back to modify answers, etc. This metric is most likely to be useful in the delayed recall task. The order of responses in this task may indicate which symbol-digit pairings the participant remembered most or least, as people tend to write responses first for the pairs they remember best. A relatively large number of boxes that were touched out-of-order may also have clinical significance.

3.2.3 Revisited boxes

We also record cells where the test-taker wrote in a box, moved on, and then came back to that box again. This could mean they didn't finish writing their answer the first time, or it could mean they're going back to change or modify their answer. Clinically this may indicate that someone is impaired enough to make mistakes, yet still cognitively healthy enough to monitor their performance and fix them.

3.2.4 Boxes left blank

We also tally the number of boxes left blank. This is a relatively rare phenomenon, but may be an indicator of impairment in visual scanning, executive function, or memory (like forgetting instructions to complete the task sequentially).

3.2.5 Timing

For each task of the dSDT, we calculate how long the test-taker spent, computing the difference in the timestamp of the first and last pen strokes in that task.[1] All tasks

[1] Note that by necessity this calculation excludes any time spent thinking prior to writing the first pen stroke.

of the dSDT use similar digit arrays, but the tasks differ in demands and amount of cognitive load.

We believe that task timing and overall timing might be indicators for cognitive status and abilities. It may also be valuable to compare the test-taker's speed in the translation task versus the copy task. The copy task will give a baseline, and will account for things like writing speed or motor dysfunction, with low cognitive load. We expect that the translation task will have a higher cognitive load, so we hope that comparing these two tasks against each other may give us a sense of cognitive function. Note that the answers to the two are by design identical, which makes the two physical tasks nearly identical.

3.2.6 Amount of Ink

We compute amount of ink as one measure of handwriting size. When the amount of ink in a cell is higher than expected for that digit relative to how much ink they use for that digit in the rest of the test, it may be a promising indicator of crossing out or overwriting.

We calculated the amount of ink in the obvious fashion, summing the Euclidean distances between each sequential sample point. We compute this for each box, as well as computing averages and standard deviations for the test as a whole, each subtest, each row, as well as each digit type.

3.2.7 Percent Ink Time

We calculate percent ink time from the percent of time the pen is in contact with paper versus not in contact. If this metric is high, we are led to believe that cognitive load was perceived low or handled easily, and the test-taker quickly knew what to write next. Conversely, if this metric is low, it indicates that the test-taker took longer to figure out answers and spent more time thinking.

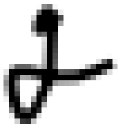

Figure 3-2: Example of a Long-Tailed Two

3.2.8 Digit height

Digit height is calculated by finding the vertical dimension of an axis-aligned bounding box around a digit, on the assumption that digits are written straight.

3.2.9 Centeredness of digit in-box

We record how centered the responses are within the response box. This may be an indicator of writing consistency, which may start to decline as cognitive status declines.

We compute centeredness in two ways. The first method uses the offset of the bounding box that encloses all of the pen strokes in that cell. The second method takes the mean of the coordinates of all of the coordinates of each stroke (i.e. the center of mass of all those strokes).

We expect the center of the bounding box and the center of mass to generally be very close to each other, but differences can arise, as for example with long-tailed two's (Figure 3-2). We record measures for both the bounding box and center of mass methods, and for each we also record a mean and standard deviation for each of the five subsections, as well as for the test as a whole.

3.2.10 Prebox pauses

Prebox pause is the interval between the first timestamp of the first stroke in a box and the last timestamp of the previous pen stroke, no matter which box it was in. We believe this pause may be a good indicator for how much time was spent thinking about whatever the test-taker plans to write next. We exclude measurements for any box for which we cannot know the true time, such as the first box after the sample

Table 3.2: Data Metrics

Total Number of Datapoints	886
Number of Unique Test-Takers	481
Number of Test-Takers with Diagnosis Information	277

section. We believe that long prebox pauses will be an indicator for cognitive decline.

3.2.11 Maximum pen speed

Maximum pen speed is a measurement of the fastest that the test-taker's pen moves when writing an answer. We believe maximum pen speed will be an interesting metric for loss of motor skills.

We find this by looking for the highest sustained speed, by creating a moving average with an empirically chosen window width of five datapoints. We record the maximum speed among all strokes within each response box. We also store the averages and standard deviations for this metric for the test as a whole, each subtest, each row, as well as each digit type.

3.3 Contextualizing Our Features

We contextualize our data to determine which metrics are better indicators of cognitive status. We initially used a visualization tool called Tableau, but found it to be too limited for our use. In particular, it had minimal ability to be programmed or automated, making it difficult to evaluate multiple datasets with the same metrics. Tableau's drag-and-drop ability is simultaneously its best feature and its most limiting feature.

We switched to a Python package called matplotlib, which is more complex, programmable, and allowed us to do complicated analyses over multiple datasets.

Chapter 4

Results

4.1 Digit Classifier Results

We trained a convolutional neural network to classify handwritten digits between 0 and 12 by expanding the MNIST dataset and using an existing model (built by the Keras team), which was designed for the original MNIST dataset [3]. We used 5-fold cross-validation with an average test accuracy of 99.008%, which is comparable to the best of the digit classifiers found online.

4.2 Visualization Results

The dataset used for all of our graphs contains only the first trial from each test-taker.

4.2.1 Population Breakdown & Data Completeness

We began with analyses that examined data completeness. These charts are helpful in highlighting imbalances in the dataset (e.g., more data from one gender than another), or missing data.

We have 149 male subjects, 145 female subjects, and 187 subjects for which gender was not recorded. We note a fairly even distribution of males versus females.

Table 4.1 shows a breakdown of our participants by handedness. We note that

Table 4.1: Population Breakdown by Handedness

Name	Count	Percent
R	411	85.0%
L	46	10.0%
Ambi R>L	1	0.0%
None	23	5.0%

Table 4.2: Population Breakdown by Diagnosis

Diagnosis	Count
Not Specified	227
Healthy Control	78
Clinical Sample	176
Parkinson's Disease	*111*
Mild Cognitive Impairment	*23*
Alzheimer's Disease	*21*
Mixed Neurological Disorders	*21*

85% of our subjects are right-handed, comparable to the 90% found in the broader adult population [18].

Table 4.2 shows our population, broken down by primary diagnosis. Some subjects have unknown diagnosis because our data came from multiple different studies, some of which were not diagnosis related.

Table 4.3 shows the data we have for cognitive screening of our subjects, consisting of combined MoCA/MMSE scores. Based on the direction of Dr. Penney, we constructed a combined MoCA/MMSE score: if there is a MoCA score, a score from the MMSE serial sevens version, or a score from the MMSE world version, we accept any of them in that order of priority. We are able to obtain slightly more data this way.

4.2.2 Revisited Boxes

Revisited boxes may indicate that the test-taker has caught mistakes or is looking back for reference or potentially other reasons. Table 4.4 shows that many subjects never revisit a box at all.

Table 4.3: Population Breakdown by Score

Score	Count - MoCA Score Only	Count - Combined MoCA/MMSE
None	379	363
11	1	1
12	4	4
13	2	2
14	3	4
15	2	2
16	1	1
17	1	1
18	3	3
19	3	3
20	6	6
21	3	3
22	9	9
23	5	5
24	9	9
25	13	13
26	7	8
27	6	9
28	6	11
29	11	12
30	7	12

Table 4.4: Number of Tests with "n" Revisited Boxes

Number of Cells Revisited	Count of Tests
0	287
1	110
2	56
3	15
4	4
5	3
6	1
7	1
8	3
9	1

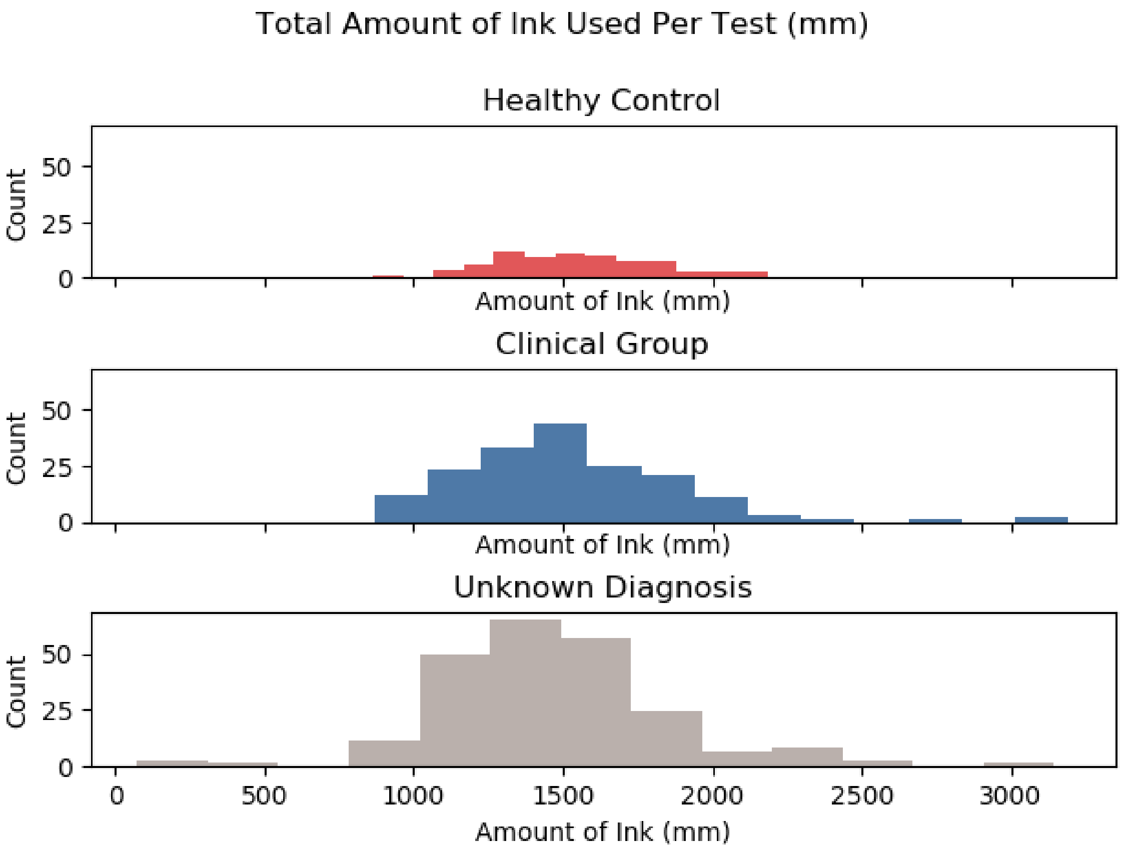

Figure 4-1: Amount of Ink Per Diagnostic Category

4.2.3 Amount of Ink

Figure 4-1 shows the amount of ink used across an entire dSDT, per diagnostic category. The Clinical group seems to use more ink than the Healthy Controls, which may be associated with cognitive impairment, potentially explained by subjects making errors and then crossing out.

Figure 4-2 compares the amount of ink in the digits 1 vs 11, for all tasks in the dSDT for each group of MoCA/MMSE scores. Each data point represents the first trial of each subject. We have colored the data, and separated it into these five charts, based on the combined MoCA/MMSE scores. The trend of the data is roughly a straight line, especially in the healthiest group, which is colored in red. This makes sense, as we expect the amount of ink in writing 11s to be roughly double the amount of ink used to write 1s. Interestingly, the only subjects whose data dips below this

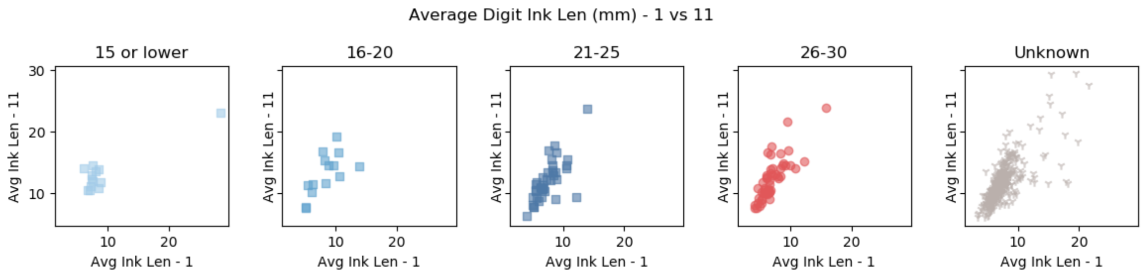

Figure 4-2: Amount of Ink for Digits 1 vs 11

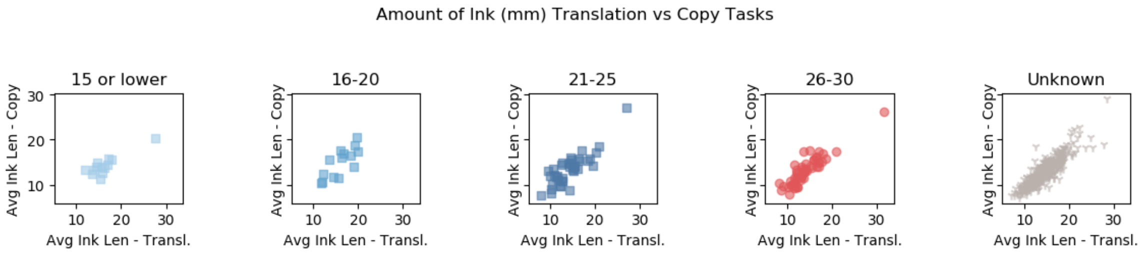

Figure 4-3: Amount of Ink for Translation vs Copy Tasks

roughly linear trend are those with cognitive deficits. This is particularly visible in the second and third plots, which represent MoCA/MMSE scores in the range 16-25.

Figure 4-3 compares the average amount of ink used per-box in the translation task compared to the amount used in the copy task. We expect these to be very similar, as the identical digits are being written. We do indeed see this similarity in the red data points, which represent our healthiest scoring group. We see a tight cluster around a roughly linear trendline. However, we do not see such a clear trendline in the groups with lower MoCA/MMSE scores. We believe this may indicate making mistakes, then later crossing them out and correcting them. This is particularly evident in the translation task, which is more cognitively demanding, in the datapoints which fall below the trendline.

Figure 4-4 displays the standard deviation of the amount of ink used per-box. We would expect to see mildly impaired patients with the highest standard deviation of amount of ink used. The reasoning is that healthy test-takers would not tend to make mistakes, and greatly impaired patients would make mistakes but not catch them. The mildly impaired patients may make mistakes, then catch and correct them. We hypothesize this would cause a higher standard deviation of amount of ink because

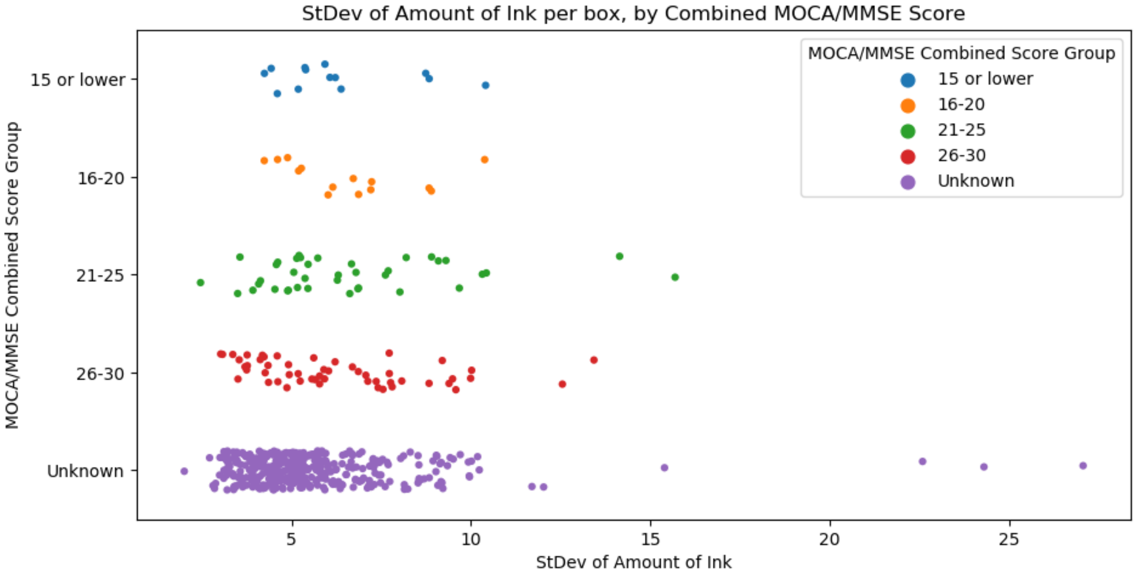

Figure 4-4: Standard Deviation of Amount of Ink

of the cross-outs and writing an updated answer.

4.2.4 Percent Ink Time

Figure 4-5 compares the percent ink time in the translation task versus the percent ink time in the copy task. We would expect a much higher percent ink time in the copy task because the task is fairly rote and has low cognitive load.

We see a very clear difference between the Healthy Control group and the Clinical Group. Among our healthy participants, we see the expected trend of a very high percent ink time in the copy task, but varied percent ink time in the translation task. We see a difference in the clinical participants, with many of them having low ink time (and presumably high thinking time), even in the copy task which should have been cognitively less intensive.

Figure 4-6 graphs the same metric, but now grouped by MoCA/MMSE scores. Surprisingly, we do not see the same deviation between groups. We see low percent ink time in the copy task in all groups – even in our healthiest group shown in red. This implies that percent ink time may be more correlated with cognitive impairment that is missed by traditional screening tests that look solely at outcome as opposed

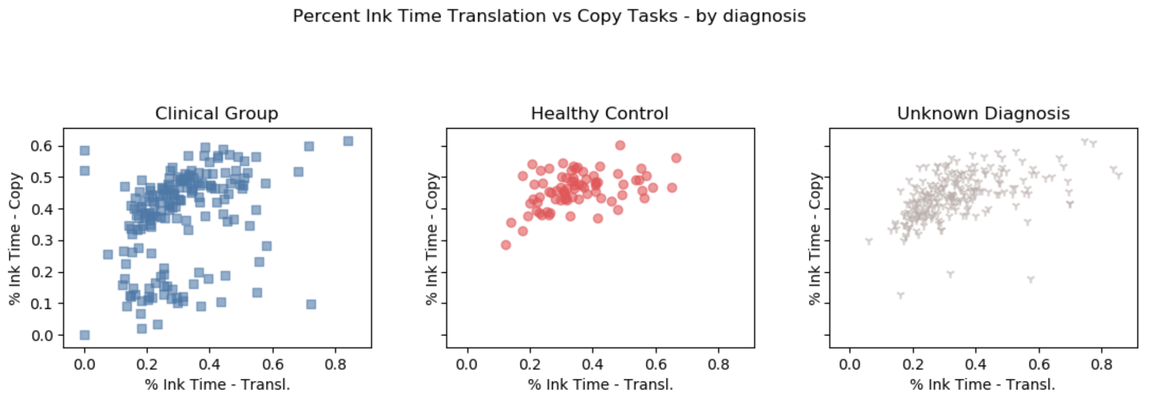

Figure 4-5: Percent Ink Time by Diagnosis

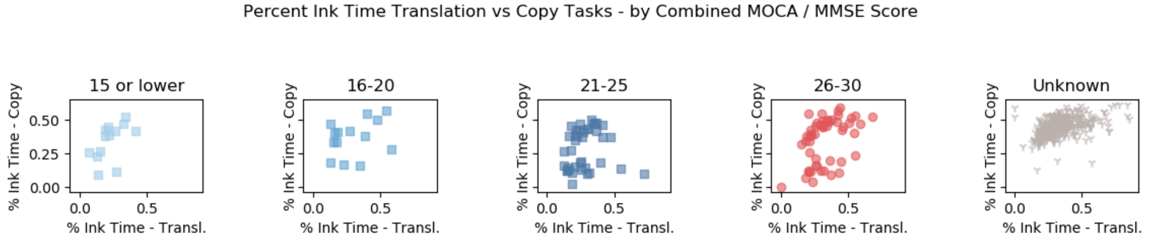

Figure 4-6: Percent Ink Time by Score

to process.

4.2.5 Digit Height

Figure 4-7 shows a scatterplot of average digit height versus total amount of ink. We expect these to be very tightly correlated. We do in fact see the healthy points in red are more tightly clustered around the trendline, with an R^2 of 0.803. We see the clinical participants less tightly grouped, with an R^2 of 0.723. In particular, we see many points that lie above the trendline, and even a few that are very far above. These points indicate a higher amount of ink than we would expect given how large their responses are on average. More ink than expected may be explained by overwriting or cross-outs, which may be correlated with mild cognitive impairment, as explained in earlier sections.

In Figure 4-8, we are looking at the standard deviation of digit height, per MoCA/MMSE score group. Interestingly, in all score groups – including our health-

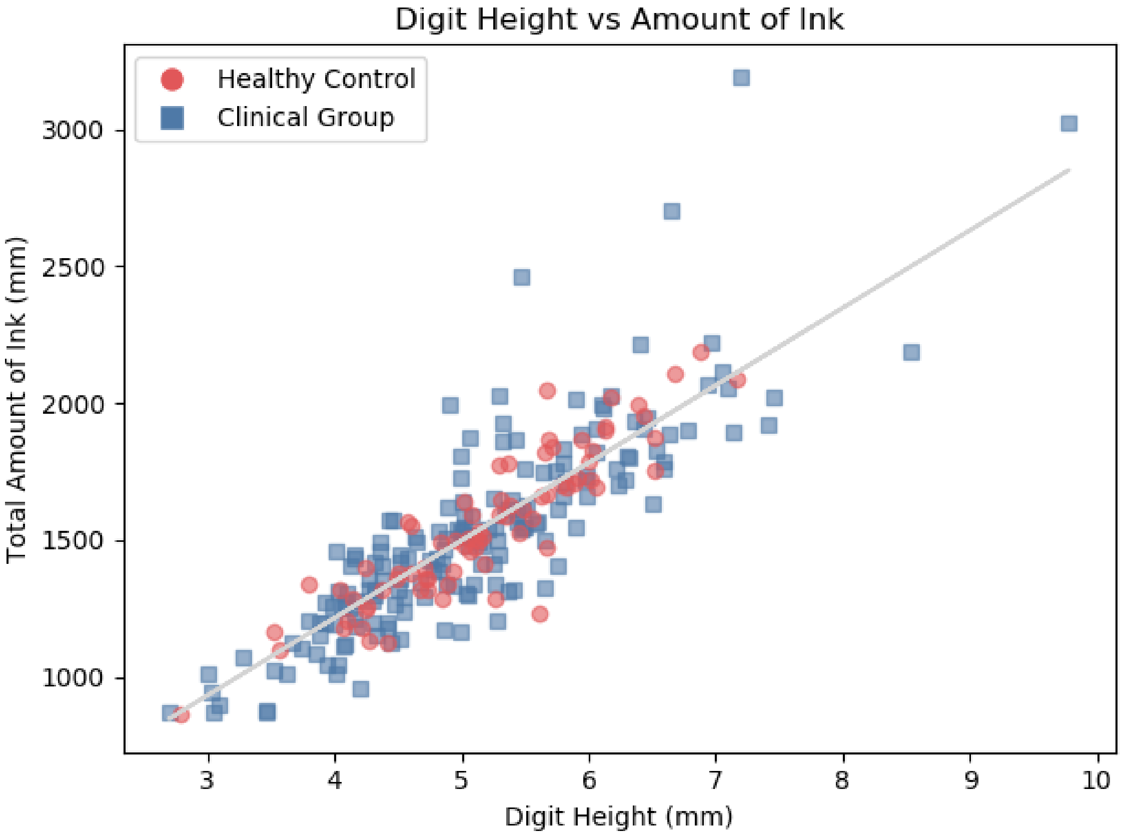

Figure 4-7: Digit Height vs Total Amount of Ink

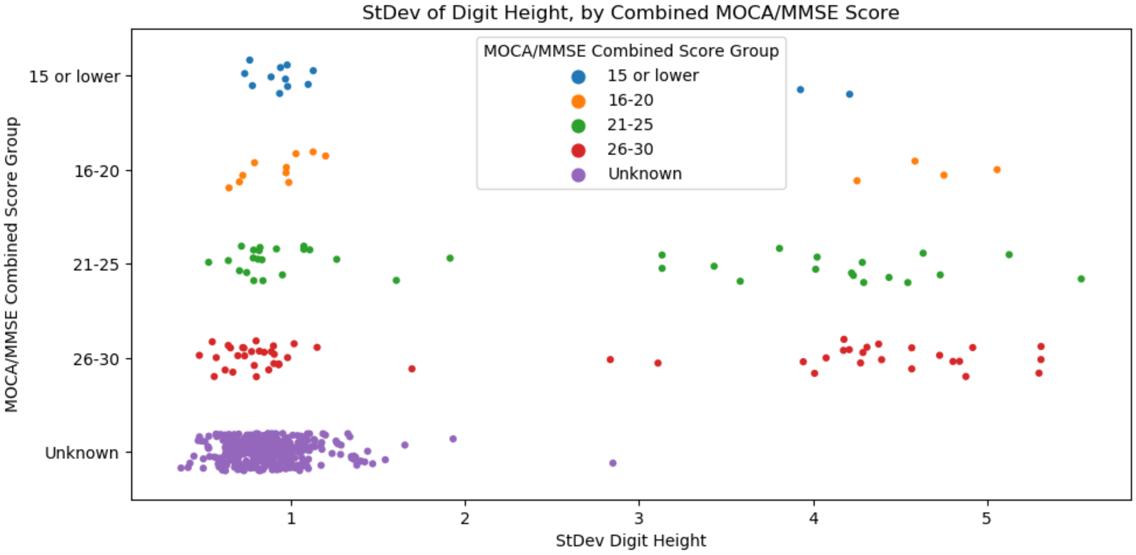

Figure 4-8: Standard Deviation of Digit Height

iest group – we notice a striation, of a group of participants with a digit height standard deviation around one, and a separate group of participants with a standard deviation of 3-6. We are not sure why this would happen, but hypothesize it might be related to our mixed clinical group, which included known movement disorders including tremor and Parkinson's Disease.

4.2.6 Centeredness of Digit in Response Box

Figures 4-9, 4-10, and 4-11 show the average off-center position of participants' responses in the response box, per task in the test. Each datapoint represents a unique participant on their first trial. We see that the delayed recall task appears to have the most laterally centered responses. All tasks tend to have a distribution where responses are written slightly above and to the left of the geometric centers of the response boxes. We have a few participants who deviate from the norm, but there does not appear to be a strong correlation with their combined MoCA/MMSE scores. We believe that off-center behavior may correlate with a combination of cognitive load and the impetus for speed, which may be interesting to explore with more data in future work.

Figure 4-12 shows the standard deviation of digit location in the x and y directions.

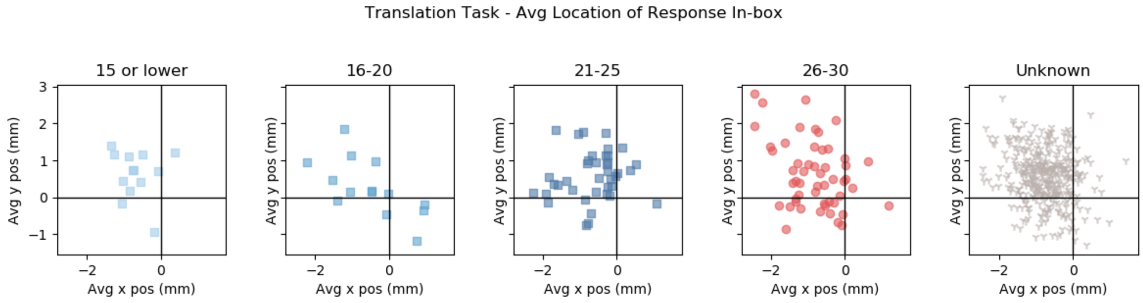

Figure 4-9: Centeredness in Box: Translation Task

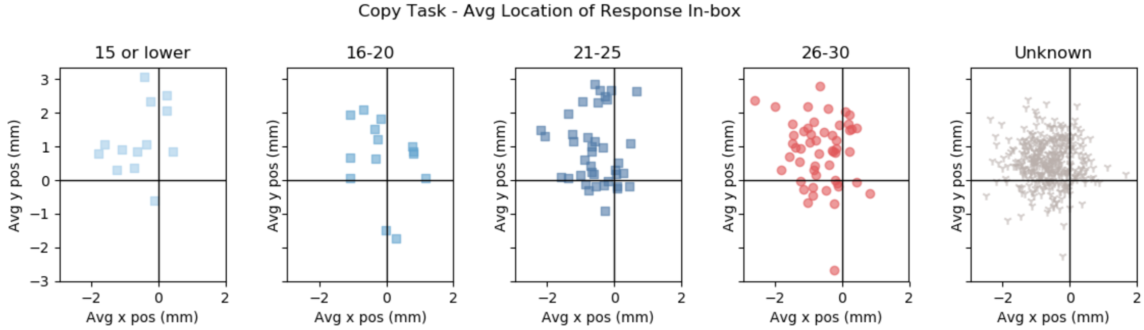

Figure 4-10: Centeredness in Box: Copy Task

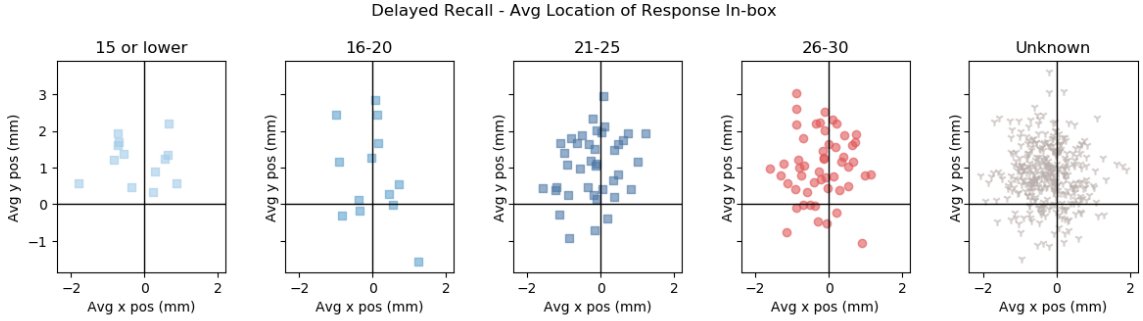

Figure 4-11: Centeredness in Box: Delayed Recall Task

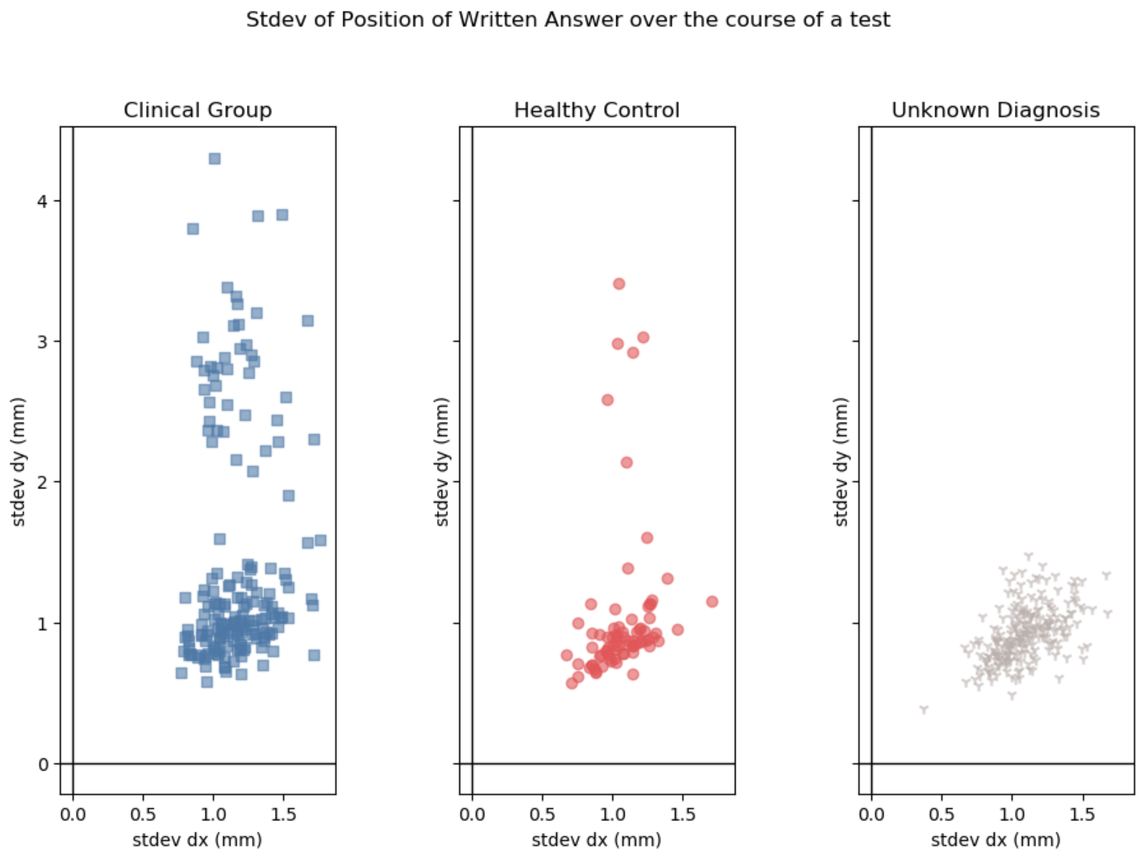

Figure 4-12: Standard Deviation of Centeredness

Interestingly, we see two clusters, of low vertical standard deviation and high vertical standard deviation. This two-cluster pattern is similar to our other graph of standard deviation in Figure 4-8. This pattern appears in both the healthy and clinical groups.

4.2.7 Centeredness of Digit in Response Box Compared with Handedness

We also analyzed handedness, as it may play a role in the location in each response box a test-taker writes their answers. Figure 4-13 shows the horizontal centeredness in response boxes as compared to handedness. Each category of handedness has its own colored box. The left vertical edge of the orange box marks the 25th percentile; the right vertical edge of the orange box marks the 75th percentile; and the vertical line in-between is the 50th percentile, and so on for each color.

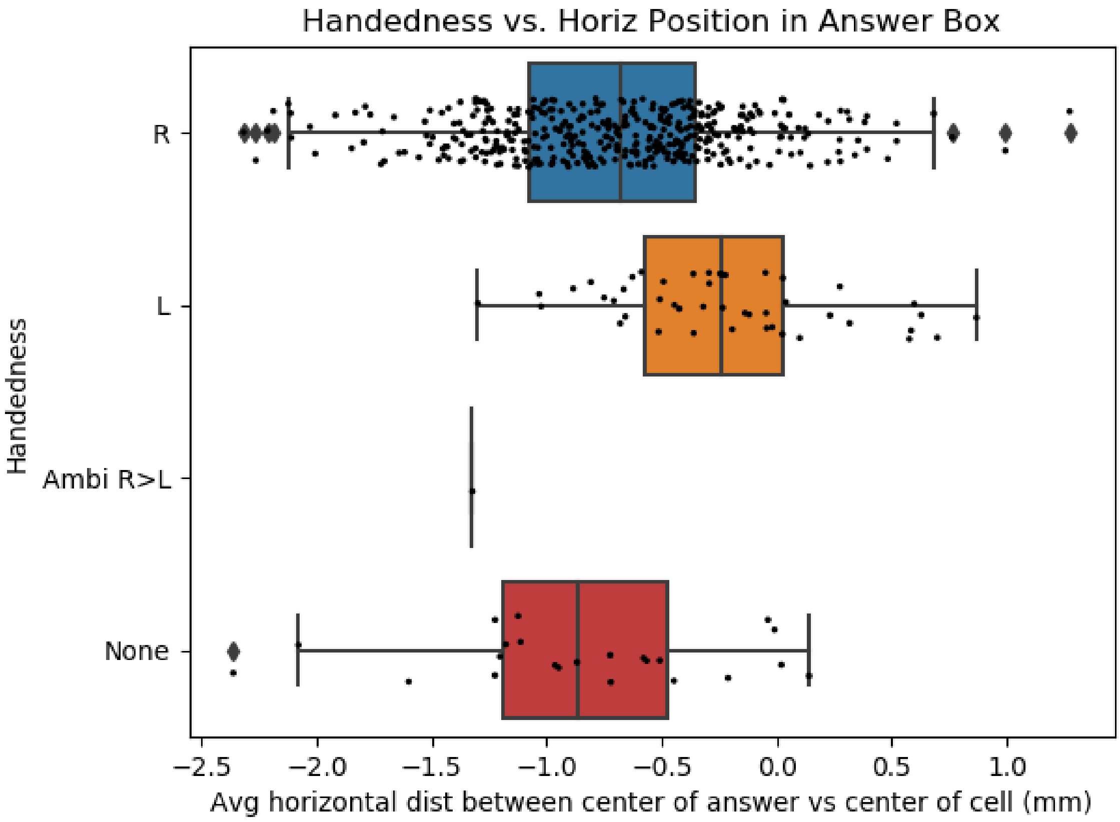

Figure 4-13: Handedness vs Horizontal Centeredness: Each colored box is composed of three vertical lines, which represent the 25th, 50th, and 75th percentile for that category of handedness.

While this is a small sample, it does seem to suggest there are differences in groups related to handedness. We performed a one-tailed t-test for two independent samples and found that the horizontal position of the answers for the right-handed participants were more skewed to the left of the response box than the left-handed participants(p = 2.88e-8).

We further subdivide Figure 4-13 by diagnostic group, to create Figure 4-14. We are hesitant to draw conclusions, as each subgroup has very little data, but the figures suggest that the trend continues within diagnostic groups, meriting further analysis.

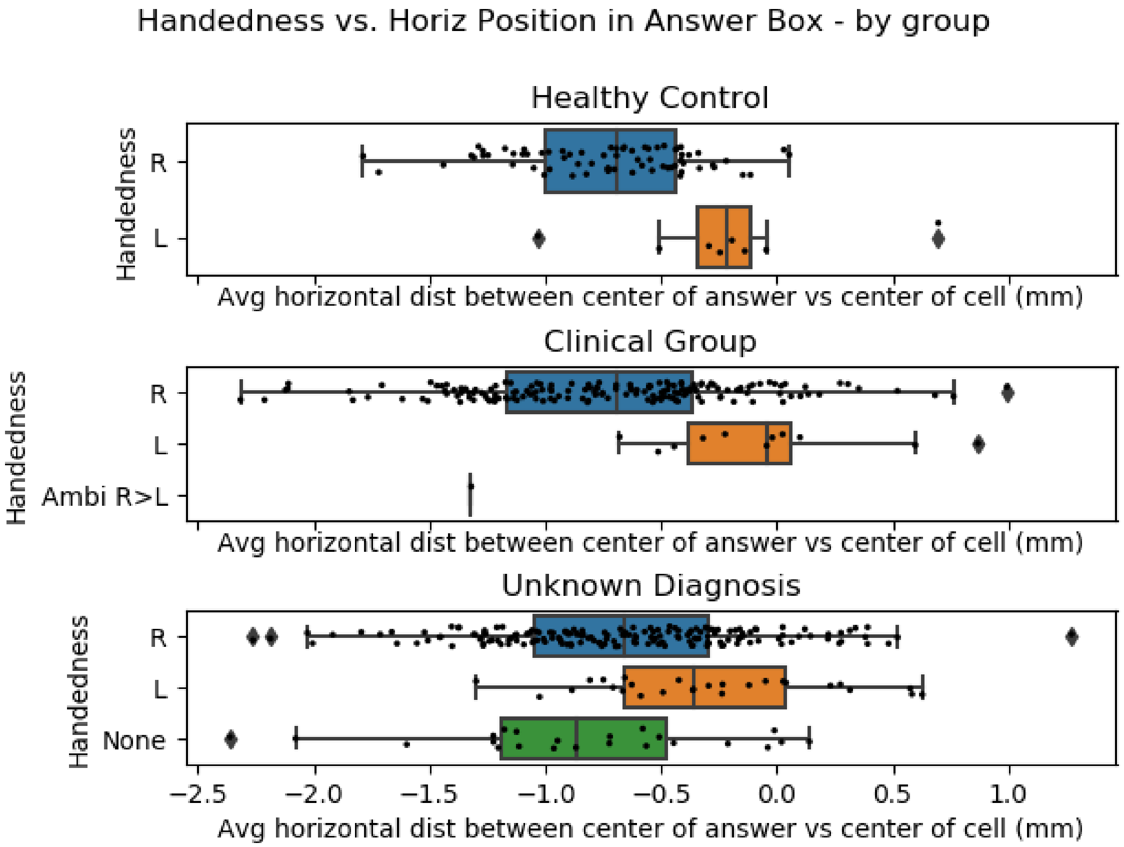

Figure 4-14: Handedness vs Horizontal Centeredness, by Diagnostic Category

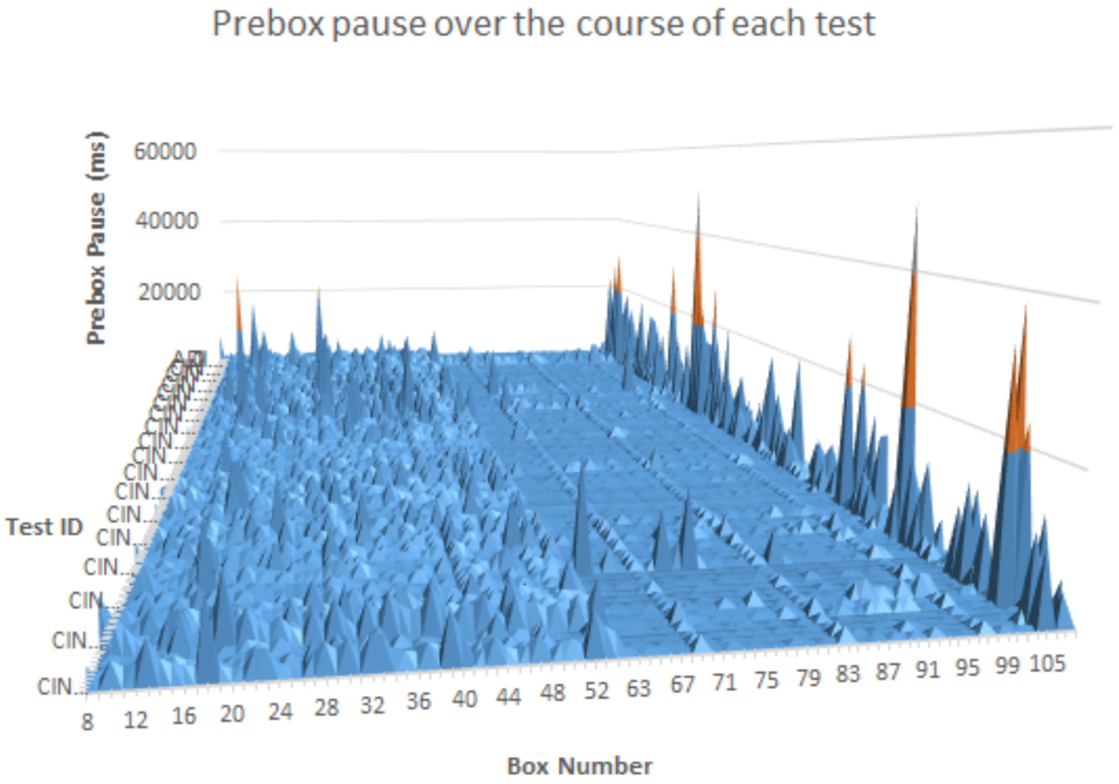

Figure 4-15: Prebox Pause 3D Surface

4.2.8 Prebox Pause

We numbered the response boxes on the test from 1 to 108 and used these as the x axis in Figure 4-15. The vertical height represents the prebox pause for each response box. Each Test ID is a different trial.

There are several interesting things here. First, there is a clear divide of the translation versus the copy tasks. That is, the left half (the translation task) has much higher prebox pauses, as we would expect, as compared to the copy task on the right half. The last 6 rows (running along the right hand side for the whole depth) have the largest spikes, which makes sense, as the delayed recall task would require more thinking time to retrieve the associations from memory. There are *plowrows* at boxes 69, 83, and 97, where a small peak consistently runs along the trials. These boxes are at the start of a new row, so it makes sense that the prebox pause would be longer, as the test-taker moves from the end of one row to the beginning of the next.

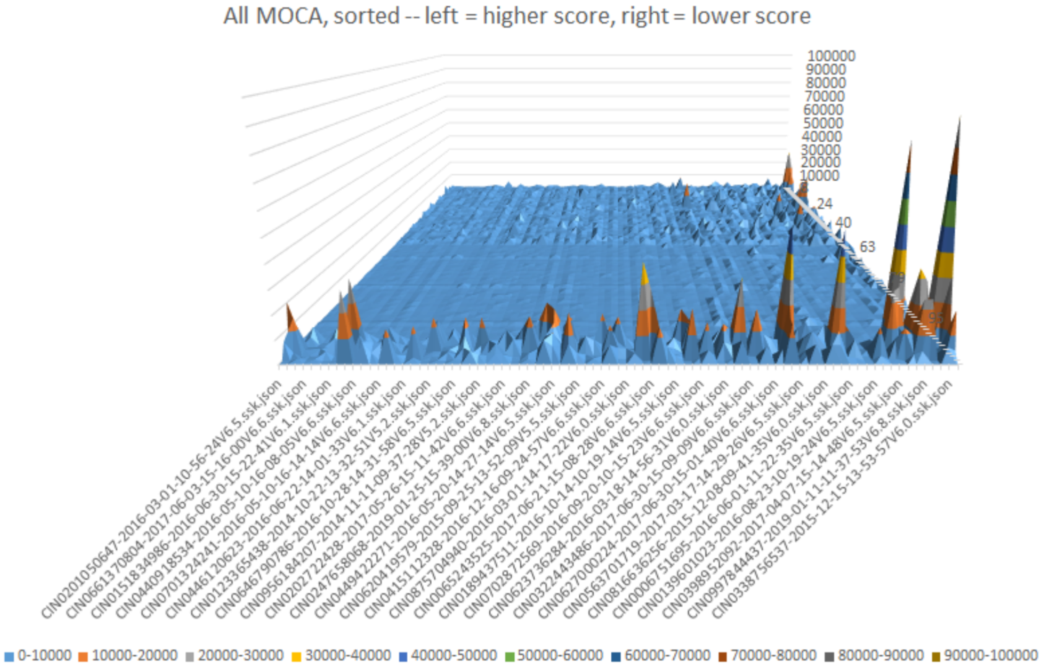

Figure 4-16: Prebox Pause 3D Surface, Sorted by MoCA/MMSE Scores

Figure 4-16 is in a similar format to Figure 4-15. It has the same axes, but is viewed from a different angle. From this perspective, the delayed recall is closest to us, and each step left or right is a different trial. A key difference in this figure is that the trials are sorted by MoCA/MMSE scores, with lower scores at the left. As we anticipated, we see the pauses get longer as we move to the right, particularly in the translation task and the delayed recall task.

Figure 4-17 shows a distribution of the prebox pause in the translation task. The data is grouped by diagnostic category. The translation task is where we would expect to see the greatest difference between clinical and healthy participants, as this task by design has higher cognitive load (and hence would also require more time to think). The colored boxes and lines represent the quartiles. For our dataset, the average prebox pause for healthy controls (green row) is almost always less than 2000 ms. However, we see many datapoints from our clinical participants (orange row) whose average prebox pause is much higher than that threshold. In fact, the median of the clinical group is above the 75th percentile in our healthy group. We believe that this can serve as a useful metric for differentiating healthy versus clinical groups.

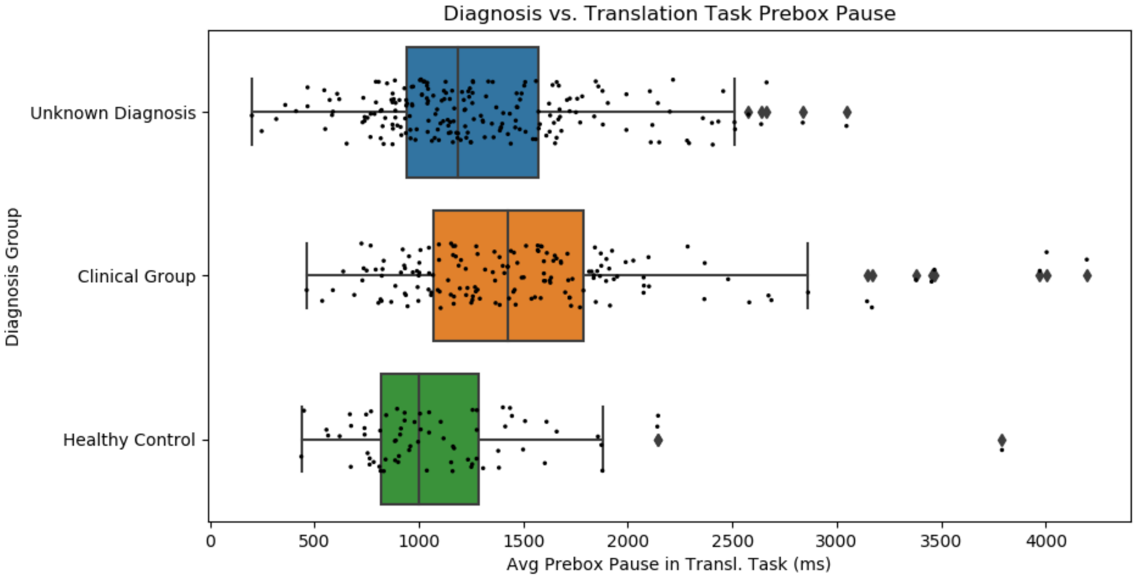

Figure 4-17: Prebox Pause in the Translation Task

Figure 4-18 shows the average prebox pause for single digit numbers versus double digit numbers. We made this comparison to see if our impaired population struggled with double digit numbers more than single digit numbers. As expected, we see a strong correlation among our healthiest population in red, where both groups of digits took roughly the same prebox pause on average. However, for our most impaired patients (at the farthest left in light blue) there does appear to be a slight difference. A few of the points are above our generally linear trend. This indicates that these patients took more time thinking before writing down double digit numbers than they did with single digit numbers. This matches our intuition that there might be a difference in cognitive difficulty between single and double digit numbers for heavily impaired participants.

Figure 4-19 and Figure 4-20 compare the prebox pauses from the translation task to those of the copy task. Figure 4-19 shows this relationship in groups by diagnosis; Figure 4-20 groups the data by MoCA/MMSE scores. We expected there might be higher average prebox pause in the translation task than in the copy task, because we would expect participants to need more time to think before writing due to the nature of the translation task. In Figure 4-19, we see a difference between the clinical versus

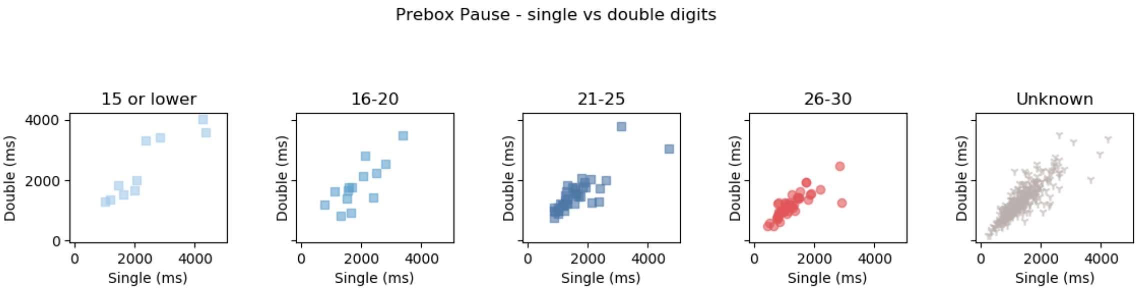

Figure 4-18: Prebox Pause in Single vs Double Digits

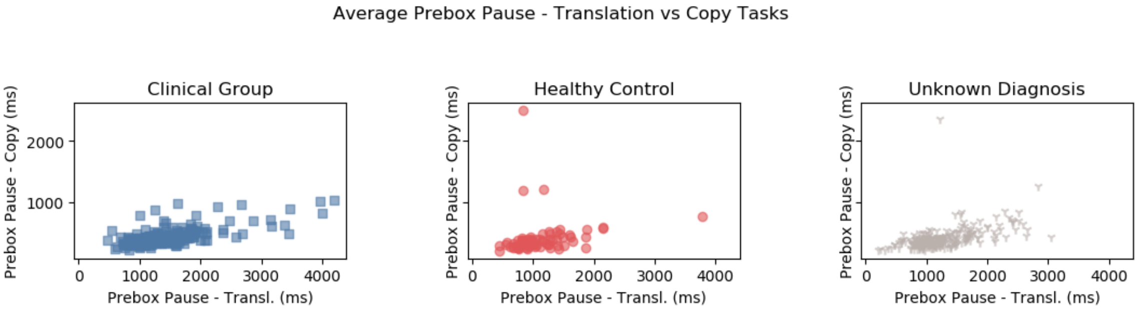

Figure 4-19: Prebox Pause in the Translation vs Copy Tasks, by Diagnosis

healthy groups. We see a large horizontal spread within the clinical group, but much more tightly clustered data in the healthy group. This highlights the same trend that we saw in Figure 4-17. Interestingly, we see a similar trend when grouping by score instead of diagnosis in Figure 4-20, with the highest ratio of translation task prebox pause averages compared to copy task prebox pause averages happening in the more impaired populations. Specifically, in all groups scoring 25 or less, we see datapoints far to the right, indicating that they needed to take much more time thinking in the translation task, as compared to their baseline from the copy task.

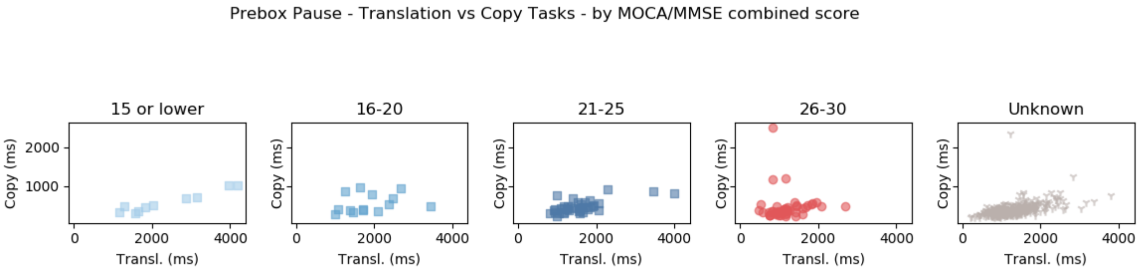

Figure 4-20: Prebox Pause in the Translation vs Copy Tasks, by MoCA/MMSE Score

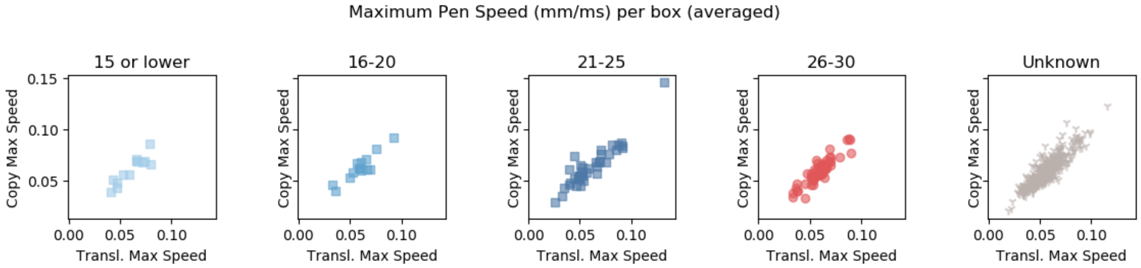

Figure 4-21: Maximum Pen Speed in the Translation vs Copy Tasks

4.2.9 Maximum Pen Speed

Figure 4-21 shows the maximum pen speed on average in the translation task versus the copy task. Generally, we see a linear trend across all panels. We see a few interesting datapoints in the 21-25 scoring group. There are a few points that lie above the linear trend. The location of these points indicates that for those participants, the maximum pen speed in the translation task is slower than we would expect, as compared to the copy task.

Chapter 5

Contributions

Our project offers three contributions. First, we have created a high-quality recognizer for the digits 0-12. Second, we have extracted a series of qualities regarding the patient's behavior while writing, which we hope can be good indicators of mental status. Finally, our graphs indicate what may be abnormal and which characteristics may be more closely tied with cognitive decline. We expect that our work will be insightful into patient condition and mental status, especially in cases of neurodegenerative diseases like Alzheimer's Disease. Since treatments for these conditions tend to be most effective early in disease progression, we expect our work may be valuable in leading towards earlier detection and a better quality of life.

Chapter 6

Future Work

The work reported so far suggests that there are a number of possible interesting directions for future work.

It would be useful to look deeper into detecting and interpreting cross-outs and overwriting. To detect, we expect that one box with high ink (compared to the mean and standard deviation of all boxes for that digit) may be an indication of cross-outs, but we have yet to define this rigorously or implement it in an algorithm.

Song's paper [21] may be useful for interpreting the final digit independent from the other ink in the box, as well as interpreting what was first written under the cross-out. Helpful literature may also be found from [1], [12], [15], [16], [23], [26].

Placeholder pen marks are small ink marks in the key or response boxes, where the subject rests their pen to hold their place. These may indicate a loss of spatial reasoning and memory, and may be an interesting feature to analyze.

It may be insightful to detect if a patient's handwriting style changes between trials that may be years apart. One idea is to use MNIST-GAN, which is able to encode style of a handwritten digit into a tiny feature vector [4].

Appendix A

Model Architecture

Architecture from Keras Team [3]

```
batch_size = 128

num_classes = 13

epochs = 12

loss = keras.losses.sparse_categorical_crossentropy

optimizer = keras.optimizers.Adadelta()


model = Sequential()

model.add(Conv2D(32, kernel_size=(3, 3),

                 activation='relu',

                 input_shape=input_shape))

model.add(Conv2D(64, (3, 3), activation='relu'))

model.add(MaxPooling2D(pool_size=(2, 2)))

model.add(Dropout(0.25))

model.add(Flatten())

model.add(Dense(128, activation='relu'))

model.add(Dropout(0.5))

model.add(Dense(num_classes, activation='softmax'))
```

Bibliography

[1] Ernest H. Beernink, Stephen P. Capps, John R. Meier, and Frederich N. Tou. Method for correcting handwriting on a pen-based computer, U.S. Patent 5,710,831, Jan. 1998. https://patents.google.com/patent/US5710831A/en.

[2] Colin D'Amore (cdamore). Object detection and classification on 2-digit mnist dataset. *GitHub repository*, 2019. https://github.com/cdamore/Object-Detection-on-2-digit-MNIST.

[3] François Chollet, Makoto Matsuyama, Stephen Merity, Junwei Pan, and Laszlo (Keras Team). Keras: Deep learning for humans: Mnist cnn. *GitHub repository*, 2018. https://github.com/keras-team/keras/blob/master/examples/mnist_cnn.py.

[4] Benjamin Bolte (codekansas). Mnist gan. *GitHub repository*, 2017. http://gandlf.bolte.cc/examples/mnist/mnist$_{g}an$/.

[5] Anne M. Damian, Sandra A. Jacobson, Joseph G. Hentz, Christine M. Belden, Holly A. Shill, Marwan N. Sabbagh, John N. Caviness, and Charles H. Adler. The montreal cognitive assessment and the mini-mental state examination as screening instruments for cognitive impairment: Item analyses and threshold scores. *Dementia and geriatric cognitive disorders*, 31(2):126–131, 2011. https://doi.org/10.1159/000323867.

[6] Li Deng. The mnist database of handwritten digit images for machine learning research [best of the web]. *IEEE Signal Processing Magazine*, 29(6):141–142. https://ieeexplore.ieee.org/abstract/document/6296535.

[7] Marshal F. Folstein, Susan E. Folstein, and Paul R. McHugh. "mini-mental state": A practical method for grading the cognitive state of patients for the clinician. *Journal of psychiatric research*, 12(3):189–198, 1975.

[8] Yassine Ghouzam. Introduction to cnn keras - acc 0.997 (top 8%). 2017. https://www.kaggle.com/yassineghouzam/introduction-to-cnn-keras-0-997-top-6.

[9] David J. Gill, Arielle Freshman, Jennifer A. Blender, and Bernard Ravina. The montreal cognitive assessment as a screening tool for cognitive impairment in parkinson's disease. *Movement disorders: official journal of the Movement Disordere Society*, 23(7):1043–1046, 2008. https://doi.org/10.1002/mds.22017.

[10] S. Hoops, S. Nazem, A. D. Siderowf, J. E. Duda, S. X. Xie, M. B. Stern, and D. Weintraub. Validity of the moca and mmse in the detection of mci and dementia in parkinson disease. *Neurology*, 73(21):1738–1745, 2009. https://doi.org/10.1212/WNL.0b013e3181c34b47.

[11] Lauren Huang. The digital symbol digit test: Screening for alzheimer's and parkinson's. Master's project, Massachusetts Institute of Technology, 2017. https://dspace.mit.edu/bitstream/handle/1721.1/122052/1108620165-MIT.pdf?sequence=1&isAllowed=y.

[12] Wolfgang Hurst, Jie Yang, and Alex Waibel. Error repair in human handwriting – an intelligent user interface for automatic on-line handwriting recognition. 1998. http://isl.anthropomatik.kit.edu/pdf/Huerst1998.pdf.

[13] Donato Impedovo and Giuseppe Pirlo. *Online Handwriting Analysis for the Assessment of Alzheimer's Disease and Parkinson's Disease: Overview and Experimental Investigation*, volume 5 of *Series on Language Processing, Pattern Recognition, and Intelligent Systems*. World Scientific, 2019.

[14] A. J. Larner. Screening utility of the montreal cognitive assessment (moca): in place of—or as well as—the mmse? *International Psychogeriatrics*, 24(3):391–396, 2012. https://doi.org/10.1017/S1041610211001839.

[15] L. Likforman-Sulem and A. Vinciarelli. Hmm-based offline recognition of handwritten words crossed out with different kinds of strokes. *The 11th International Conference on Frontiers in Handwriting Recognition*, 2008. http://eprints.gla.ac.uk/59027/1/id59027.pdf.

[16] Zhouchen Lin. Cleaning up of handwriting intra-stroke and inter-stroke overtracing, U.S. Patent 7,567,711 B2, Jul. 2009. https://patents.google.com/patent/US7567711B2/en.

[17] Tom Y. Ouyang and Randall Davis. A visual approach to sketched symbol recognition. *Proceedings of the International Joint Conference on Artificial Intelligence*, 2009. http://rationale.csail.mit.edu/publications/Ouyang2009IJCAI.pdf.

[18] Sara M. Scharoun and Pamela J. Bryden. Hand preference, performance abilities, and hand selection in children. *Frontiers in Psychology*, 5(82), 2014. https://www.ncbi.nlm.nih.gov/pmc/articles/PMC3927078/.

[19] Douglas W. Scharre. Preclinical, prodromal, and dementia stages of alzheimer's disease. *Practical Neurology*, pages 36–47, June 2019. https://practicalneurology.com/articles/2019-june/preclinical-prodromal-and-dementia-stages-ofalzheimers-disease/pdf.

[20] Shao-Hua Sun (shaohua0116). Multi-digit mnist for few-shot learning. *GitHub repository*, 2019. https://github.com/shaohua0116/MultiDigitMNIST.

[21] Yale Song, Randall Davis, Kaichen Ma, and Dana L. Penney. Balancing appearance and context in sketch interpretation. *Proceeding of the International Joint Conference on AI*, 2016. http://groups.csail.mit.edu/mug/pubs/SongDavisMaPenneyIJCAI16.pdf.

[22] Greg Surma. Digit recognizer - introduction to kaggle competitions: Solving mnist digit recognition task (0.995). 2018. https://towardsdatascience.com/digit-recognizer-introduction-to-kaggle-competitions-with-image-classification-task-0-995-268fa2b90e13.

[23] Diar Tuganbaev and Dmitri Deriaguine. Method of stricken-out character recognition in handwritten text, U.S. Patent 8,472,719 B2, Jun. 2013.

[24] Judie Walton. Handwriting changes due to aging and parkinson's syndrome. *Forensic Science International*, 88(3):197–214, August 1997. https://www.sciencedirect.com/science/article/pii/S0379073897001059.

[25] Xuan Yang and Jing Pu. Mdig: Multi-digit recognition using convolutional nerual network on mobile. 2015. https://www.semanticscholar.org/paper/MDig-%3A-Multi-digit-Recognition-using-Convolutional-Yang/76c44858b1a3f3add903a992f66b71f5cdcd18e3.

[26] Jin-Yong Yoo, Min-Ki Kim, Sang Yong Ban, and Young-Bin Kwon. Line removal and restoration of handwritten characters on the form documents. *IEEE*, 1997. https://ieeexplore.ieee.org/document/619827.

[27] Josh Zeigler. How big is too big for json? 2012. https://joshzeigler.com/technology/web-development/how-big-is-too-big-for-json.