2014

Statistical analysis of the variability and reliability of eye-tracking test in measuring mild traumatic brain injury

Xi He

Statistical Analysis of the Variability and Reliability of Eye-Tracking Test in Measuring Mild Traumatic Brain Injury

By

Xi He

Honors Thesis

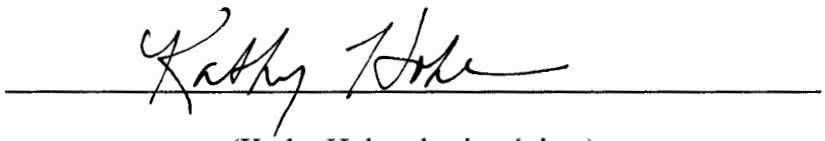
In

Department of Mathematics and Computer Science
University of Richmond
Richmond, VA

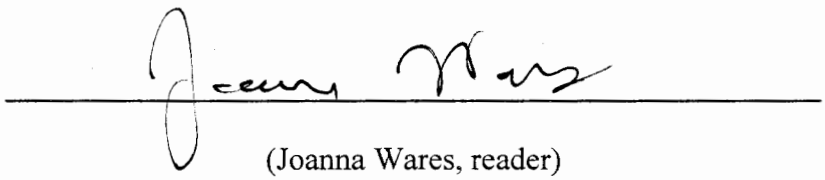
May 2, 2014

Advisor: Dr. Katherine W. Hoke

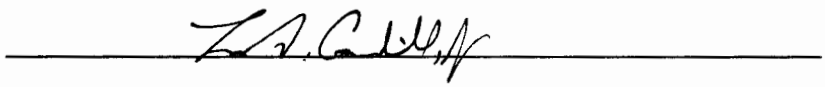
The signatures below, by the thesis advisor, a departmental reader, and the honors coordinator for mathematics, certify that this thesis, prepared by Leonhard Euler, has been approved, as to style and content.



(Kathy Hoke, thesis advisor)



(Joanna Wares, reader)



(Lester Caudill, honors coordinator)

Abstract

Saccadic eye-tracking tests have been advocated as a useful tool to distinguish mTBI patients from healthy people. However, intra-individual variances sometimes interfere with the interpretation of eye-tracking results, especially in experiments when group size is restricted. This study analyzes eye-tracking results of 14 mTBI patients taking the test twice with no medical administration in between. Using more accurate models to fit each individual's result, variables such as asymptote (of the fit functions) and hypothetical values for peak velocity, peak acceleration, and duration are derived for variability analysis. We conclude that the asymptotes for peak velocity and peak acceleration are the most reliable variables for future experiments to study, in that these variables have the highest intraclass correlation coefficient and confidence intervals. Moreover, predicted values require fewer participants in each group for the experiment to detect statistical differences between the experimental group and control group. Whichever variable future studies choose to examine, we recommend at least one replication of the same test to be conducted.

1 Introduction

Eye tracking is the procedure of measuring the motion of one's point of gaze. It has been advocated as an objective assessment of the brain following mTBI (mild Traumatic Brain Injury) that has shown promise as a "user friendly, low cost, non-invasive definitive approach" [10] (also see [4] and [8]). Recent research by Cifu et al. [4] supports this view by suggesting that mTBI subjects track moving targets less accurately than normal subjects, and that eye movement differences between the two groups can be detected and quantified.

However, results from eye-tracking tests could be prone to measurement error stemming from intra-individual variability. Intra-individual variability is the idea that every individual at a given time is a complicated "configuration of characteristics" [9]. While some of these characteristics are relatively stable, others are constantly changing. Hence, when the same person takes the same eye tracking test twice with no external intervention in between, if the person has a high intra-individual variance, the two test results may mislead us to conclude that the test is taken by two different people. Bollen et al. [2] find that variability within individuals affects the interpretation of repeated eye-tracking tests taken before and after medical treatment. In addition, substantial intra-individual variability decreases the likelihood of detecting statistically significant differences between mTBI patients and normal subjects, particularly when group size is small.

This study contributes to the research of Cifu et al. [4] by analyzing intra-individual variability in their data and exploring the reliability of eye-tracking tests. We examine test results of a group of mTBI patients without medical administration, who take the eye-tracking test twice, to identify the amount of variance that is attributed to intra-individual variability in results from the two times the test is taken. Our goal is to try to answer the question: can eye-tracking tests be used to distinguish between healthy people and people with brain injury?

We use de-identified data collected from 61 active-duty veterans who have been diagnosed with mTBI. The data come from an experiment [4] in which the effects of treatment with a hyperbaric chamber were explored. The 61 participants were randomly assigned to breathe one of three oxygen mediums in the hyperbaric chamber at 2.0 ATA, specifically 10.5%, 75%, or 100% oxygen. The sham-control (10.5% oxygen at 2.0 ATA) simulated a placebo or sham exposure. Participants took eye-tracking tests twice, once (baseline) just before treatment (time A), and once just after treatment (time B).

Specifically, the eye tracking data used in this paper were collected in the following manner (by experts in this field). We quote Cifu et al. [4]:

> *Horizontal and vertical binocular gaze data, at 500 samples per second, were recorded using a head mounted video-based binocular eye tracker (Eyelink II, SR Research, Kanata, Ontario, CAN). The subject's head was supported by an adjustable chin rest cup in order to minimize head movement. Stimuli covering $\pm 20^{\circ}$ horizontally and $\pm 13^{\circ}$ vertically were presented at 120 Hz on a 24-in LCD monitor placed 75 cm from the subject's eyes in a darkened room. The monitor display's height was adjusted with the center of the screen corresponding to the*

center of the pupillary plane. Before recording commenced, calibration and validation of the eye tracker was immediately performed at three points along each cardinal axis. The target stimulus was a white annulus, sized to occupy 0.25° of visual angle, with a high-contrast center point of 0.1° presented on a black background. Stimuli consisted of random, unpredictable step target movements in both the horizontal and vertical directions. To prevent fatigue, subjects were allowed to close their eyes and rest between each recording.

A several step process was used to analyze eye position data....During automated analysis, the criteria for detecting a saccade required that the amplitude of the movement was greater than ±0.1°, the duration of the saccade fell within a predetermined minimum and maximum time limit, and that the calculated velocity and acceleration values (based on a two-point central difference method) were greater than ±20°/s and 400°/s2, respectively, but also did not exceed a set of predetermined upper limits (in absolute value) for both velocity and acceleration. For any saccadic eye movement, the time, location, and amplitude of the saccade, as well as, its direction, duration, peak velocity, and peak acceleration and deceleration reached during the movement were determined and stored in a measurement summary file for later statistical analysis.

Assessments of eye tracking include the measurement of saccades, smooth pursuit eye movement, and fixation (see [3] and [6]). We focus on saccades, which refer to rapid shifts of eye fixation [7]. Measurements such as amplitude, peak velocity, peak acceleration, and duration are used to describe saccades. Amplitude (or magnitude, position) is the size of the eye movement, usually measured in degrees or minutes of arcs [11] (our study uses degrees). Peak velocity is the highest speed during the saccade, and peak acceleration is the highest acceleration during the movement. Duration is the amount of time it takes to complete the saccade. Figure 1, 2, and 3 are graphical illustrations of these variables.

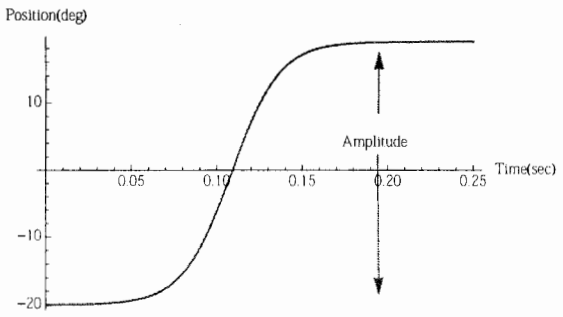


Figure 1: Position and amplitude. Velocity is the rate of change of position and the slope of the position function. The largest slope (at around 0.11 sec) corresponds to peak velocity.

When studying saccades, we look at the relationships between peak velocity and amplitude, peak acceleration and amplitude, and duration and amplitude. These relationships are called the main sequence data. Agreement has not been achieved on the exact definition

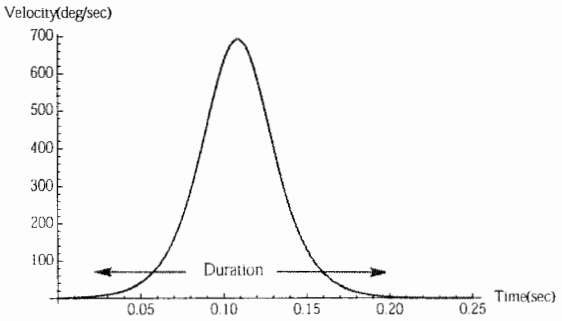


Figure 2: Velocity. Peak velocity is the absolute value of the highest velocity. Duration is the time to complete the saccade, that is, from when velocity is 0 to 0 again.

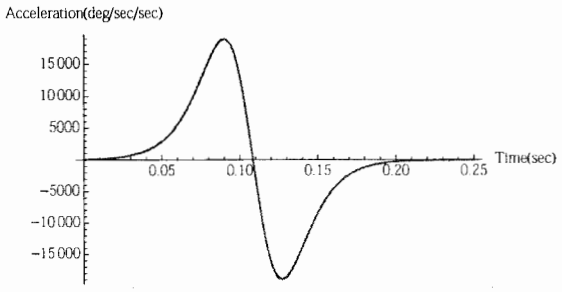


Figure 3: Acceleration. Peak acceleration is the absolute value of the highest acceleration.

of main sequence. Leigh and Zee [7] and Bollen et al. [2] use the term exclusively for the relationship between amplitude and peak velocity, while researchers at the University of Liverpool [11] define it more broadly to include relationships between amplitude and duration as well. Bahill, Clark, and Stark [1] also concur with the latter definition. This study chooses to use the definition indicated by Bollen et al. [2] in order to make a reference to their study.

2 Model

To determine whether our eye-tracking test is a reliable measurement to detect differences between patients and controls, we consider test results of each participant in the sham group at time A and B, since the sham group is the only group to complete replication tests without oxygen treatment. Specifically, we first decide on the optimal regression models describing the relationships of absolute value of amplitude with absolute value of peak velocity, peak acceleration, and duration. (Duration is positive while the other measurements are not always positive, since they denote directions. From now on, we talk about all measurements in terms of absolute value.) Then we derive from these models variables including asymptotes and hypothetical peak velocities, peak accelerations, and durations at amplitude 1° and 5° (more specific definitions will be given later). These variables will be used in the variability analysis in later sections.

2.1 Peak velocity and amplitude

We first examine the relationship between each sham participant's peak velocity and amplitude. There are a total of 21 sham patients, but only 14 of them have meaningful paired data of peak velocity and amplitude at both time A and B. Three participants have fewer than four paired data at time A, three participants have results at time B but not time A, and one participant has only one paired data at time B, so the results of these seven participants are not used.

Two regression models have been used in previous literature to fit the relationship between peak velocity and amplitude. We fit both models with our data and use the coefficient of determination R^2 as a major criteria to decide which model has a closer fit. $R^2 \in [0, 1]$, and generally speaking, the larger R^2, the better the fit. The first model is an exponential function given in *The Neurology of Eye Movement* by Leigh and Zee [7] and used by Cifu et al. [4] to account for the relationship between peak velocity and amplitude for all mTBI patients. Although the approach taken by Leigh and Zee [7] and Cifu et al. [4] is to combine all saccades from multiple individuals and fit the main sequence curve to this data, and in this study only results of the sham group are examined, the model has proven to be a closer fit than most other known models. So we test this model with our data. The function is:

$$\text{Peak Velocity} = V_{max} \cdot (1 - e^{-\text{Amplitude}/C}), \tag{1}$$

where V_{max} and C are constants. V_{max} is the asymptotic peak velocity, and C defines the exponential rise.

Following a similar approach as Leigh and Zee [7], we combine all saccades from all 14 patients and use function (1) to fit our data in Mathematica using the NonlinearModelFit command, which finds a least-squares fit. We get the following fit for time A:

$$\text{Peak Velocity} = 494.557 \cdot (1 - e^{-0.128 \cdot \text{Amplitude}}),$$

with $R^2 = 0.933$. For time B:

$$\text{PeakVelocity} = 518.404 \cdot (1 - e^{-0.130 \cdot \text{Amplitude}}),$$

with $R^2 = 0.937$. See Figure 4 and 5 for graphical demonstrations of the two fits.

Another model that describes the relationship between peak velocity and amplitude that is used by Bollen et al. [2] is:

$$\text{Peak Velocity} = A \cdot \text{Amplitude}^B, \tag{2}$$

where A and B are constants.

Using function (2) to fit our data, for time A we get :

$$\text{Peak Velocity} = 83.965 \cdot \text{Amplitude}^{0.563},$$

with $R^2 = 0.927$. For time B:

$$\text{Peak Velocity} = 86.085 \cdot \text{Amplitude}^{0.573},$$

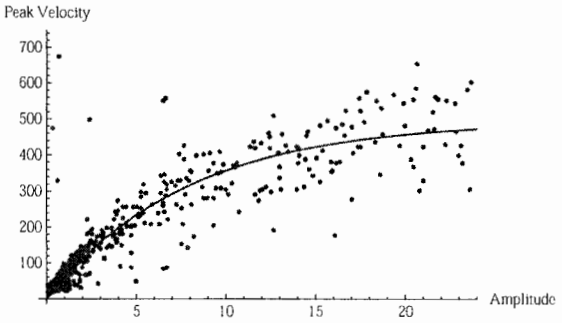


Figure 4: The regression function Peak Velocity $= 494.557 \cdot (1 - e^{-0.128 \cdot \text{Amplitude}})$ and all peak velocities vs. amplitude for sham at time A.

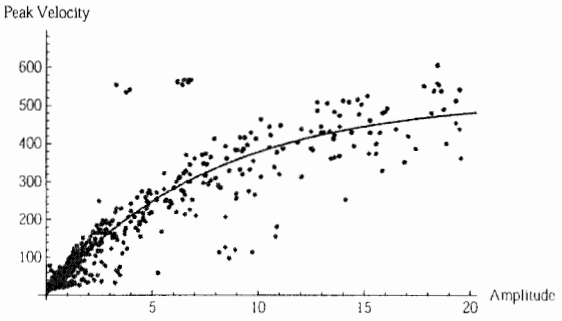


Figure 5: The regression function Peak Velocity $= 518.404 \cdot (1 - e^{-0.130 \cdot \text{Amplitude}})$ and all peak velocities vs. amplitude for sham at time B.

with $R^2 = 0.922$. See Figure 6 and 7 for fittings at the two times using model (2).

Comparing the R^2 values of regression models for all 14 sham patients, model (1) is a better fit. Next we compare the two models for fitting individual data. The results are summarized in the table below.

92.86% of the individual fits using model (1) at both time A and time B have higher R^2 values than model (2). So we proceed with function (1) as our model for peak velocity and amplitude and use it to compute asymptote and hypothetical peak velocity at amplitudes 1° and 5° for each individual for later analysis. Asymptote refers to the V_{max} coefficient in function (1), and hypothetical peak velocity at amplitude n° is the velocity calculated by setting amplitude to n in the function.

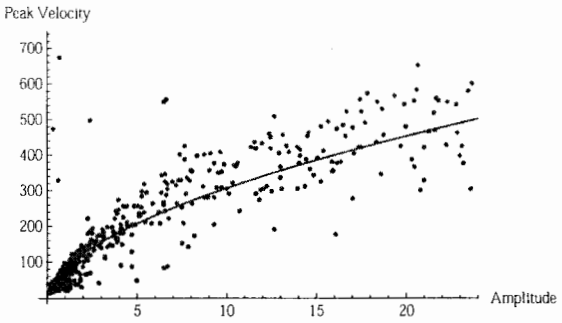


Figure 6: The regression function Peak Velocity $= 83.965 \cdot \text{Amplitude}^{0.563}$ and all peak velocities vs. amplitude for sham at time A.

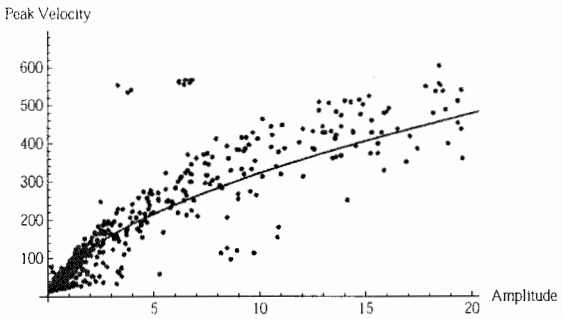


Figure 7: The regression function Peak Velocity $= 86.085 \cdot \text{Amplitude}^{0.573}$ and all peak velocities vs. amplitude for sham at time B.

	time A		time B	
	model (1)	model (2)	model (1)	model (2)
1	0.978	0.953	0.910	0.853
2	0.992	0.964	0.994	0.964
3	0.988	0.935	0.924	0.843
4	0.984	0.965	0.851	0.788
5	0.908	0.903	0.840	0.950
6	0.865	0.842	0.870	0.773
7	0.926	0.857	0.990	0.943
8	0.952	0.905	0.979	0.940
9	0.985	0.950	0.995	0.966
10	0.908	0.903	0.970	0.922
11	0.961	0.934	0.991	0.969
12	0.877	0.882	0.990	0.952
13	0.991	0.928	0.987	0.920
14	0.929	0.907	0.979	0.937

2.2 Peak acceleration and amplitude

The relationship between peak acceleration and amplitude is usually modeled by function (1) by changing peak velocity in the function to peak acceleration. The graphs for the fits are figure 8 and 9.

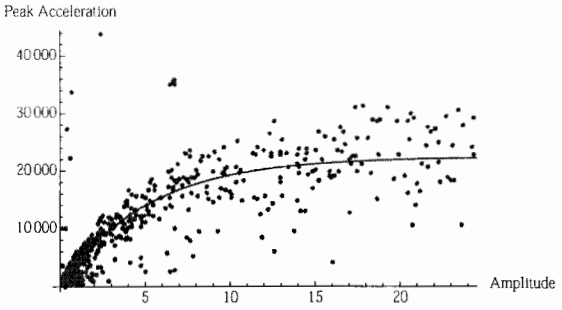


Figure 8: The regression function Peak Acceleration $= 22427.6 \cdot (1 - e^{-0.196 \cdot \text{Amplitude}})$ and all peak acceleration vs. amplitude for sham at time A. $R^2 = 0.893$.

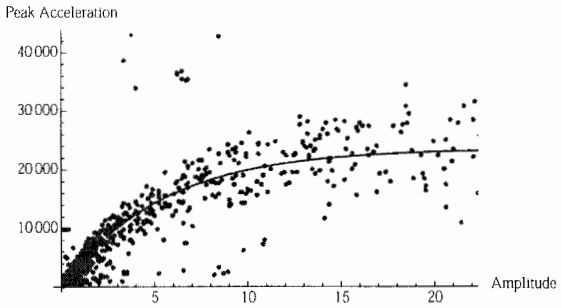


Figure 9: The regression function Peak Acceleration $= 23547.6 \cdot (1 - e^{-0.188 \cdot \text{Amplitude}})$ and all peak acceleration vs. amplitude for sham at time B. $R^2 = 0.888$.

2.3 Duration and amplitude

The regression model for duration and amplitude used in *The Neurology of Eye Movement* [7] is in the form of function (2) except peak velocity is changed to duration. This function has been the most widely used one as well. The fittings for our data at time A and time B are Figure 10 and 11.

We fit function (2) for each individual and get asymptote and hypothetical duration. The asymptote for duration refers to the A coefficient in function (2).

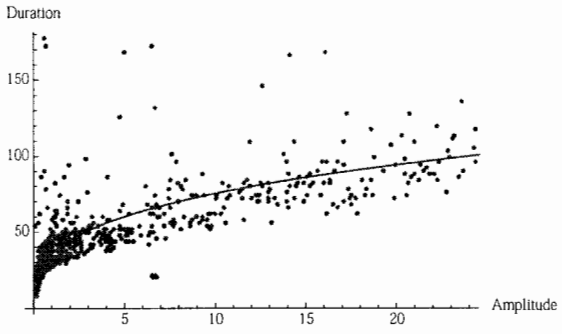


Figure 10: The regression function Duration $= 35.627 \cdot \text{Amplitude}^{0.325}$ and all duration vs. amplitude for sham at time A. $R^2 = 0.923$.

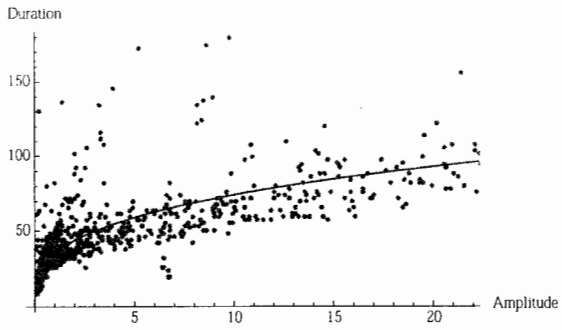


Figure 11: The regression function Duration $= 35.0386 \cdot \text{Amplitude}^{0.0.326}$ and all duration vs. amplitude for sham at time B. $R^2 = 0.917$.

3 Variability analysis

We will use asymptotes and calculated values at 1° and 5° for peak velocity, peak acceleration, and duration to explore the usefulness of our variables in detecting differences between patients and controls. We follow the procedure in Bollen et al. [2].

For the purpose of the study, we conduct analysis in the following steps:

1. We first perform statistical tests to compare variables at time A to the variables at time B. For example, we compare all the individual asymptotes of peak velocity of the sham patients at time A with time B. First, we test if the variables (at both times) meet the assumptions for a parametric test. If so, we use the Student's t-test of paired differences to see if the mean difference between variables is equal to 0. If the assumptions are not met, we use a nonparametric test with the same functionality: the Wilcoxon signed rank test.
2. Next we calculate the correlation coefficient r (Pearson's r) between variables at time A and variables at time B. The correlation coefficient is a measure of the strength of the linear correlation between two variables. It ranges from -1 to 1, and the higher the linear correlation, the closer r is to 1.

3. Then we estimate the intraclass correlation coefficient R_{icc}. The R_{icc} is a measure of the share of between-individual variance in the total variance between two variables, given by the formula described by Fleiss [5]:

$$R_{icc} = \frac{\sigma_T^2}{\sigma_T^2 + \sigma_e^2}, \tag{3}$$

where σ_T^2 is the variance for the error-free score T, and σ_e^2 is the variance for random error e. R_{icc} describes how strongly measurements for each individual resemble each other. An R_{icc} of 0 indicates that the measurements are unreliable, and that differences between patients are due exclusively to random measurement error. An R_{icc} of 1 means there is no measurement error. We also compute a confidence interval, an interval estimate, for R_{icc}.

4. Finally, we use results from (3). to determine the minimum number of participants n required to detect a difference of δ between patients and controls when they take a single measurement, double measurements, and beyond.

In the next section, we illustrate and elaborate on this analysis using predicted peak velocity at amplitude 1° as an example.

4 Analysis for predicted peak velocity at amplitude 1°

4.1 Mean

A common test to compare two sets of variables of the same individual is the Student's t-test of paired differences. It is a parametric test used to determine if the means of two populations are significantly different from one another, given that both sets of data follow normal distributions. Hence, before proceeding to the Student's t-test, we test the normality of the data.

In R, we generate the quantile-quantile plot (Q-Q plot) for the two sets of data at time A and time B (Figure 12 and 13). Q-Q plot is a probability plot used to visually check if data follow normal distributions. Points from the data are plotted with a 45° reference line. The closer the points follow the trend of the reference line, the higher likelihood that the data are normally distributed. Visually, peak velocities of sham participants at 1° at time A and time B appear to be approximately normally distributed. We also numerically test the normality of the data using the Shapiro-Wilk normality test. For time A, we get a p-value of 0.680, and for time B: 0.839, both much higher than 0.05. This implies that our two sets of data are both approximately normally distributed. Hence, based on our visual and numerical evidence, we can assume that both sets of data follow normal distributions.

With the assumption of normality met, we can use the parametric paired Student's t-test to decide whether the sham patients' predicted peak velocities at 1° at times A are significantly different from those at time B. We use the t.test() command (with the option

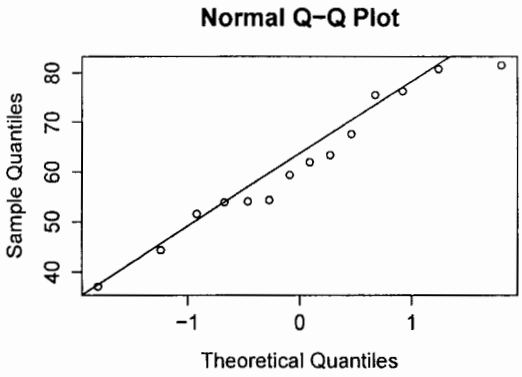


Figure 12: Q-Q plot of calculated peak velocity at 1° in time A

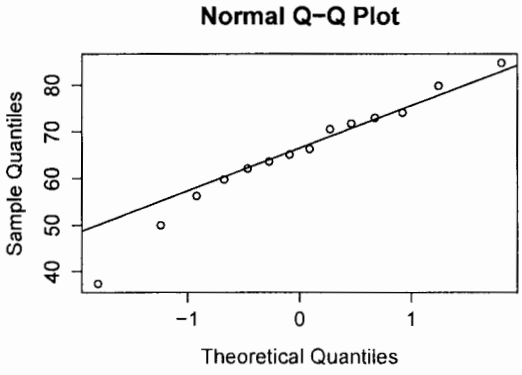


Figure 13: Q-Q plot of calculated peak velocity at 1° in time B

"paired" set to true) in R, which gives a p-value of 0.417, larger than 0.05. This implies that we are unable to reject the null hypothesis that the true difference in means for the two data is zero. Thus, there is no reason to believe the means are different.

Bollen et al. [2] use the nonparametric Wilcoxon signed rank test, which serves the same purpose as the Student's t-test but does not assume normality of the data. To make a comparison with their study, we test our data with this test as well. The Wilcoxon signed rank test gives a p-value of 0.308, larger than 0.05. Hence, we are unable to reject the null hypothesis that the true location shift in the two data is not equal to zero.

Both the parametric paired Student's t-test and the nonparametric Wilcoxon signed rank test reach the same conclusion that the results from the test A and test B have approximately the same mean.

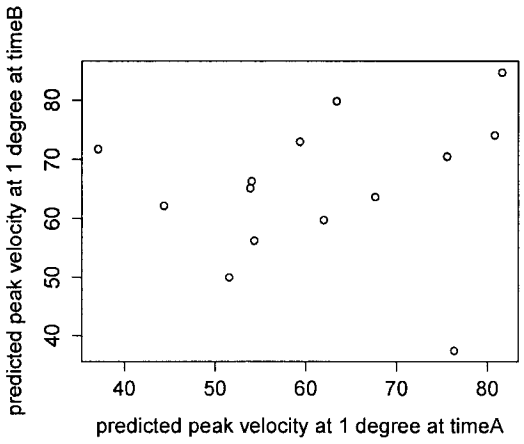


Figure 14: The correlation between the predicted peak velocities at 1° in time A and time B is low. $r = 0.152$

4.2 Correlation Coefficient

The correlation between calculated peak velocities at 1° in times A and time B can be represented by the correlation coefficient r. It is defined as the covariance of two data sets divided by the product of their individual standard deviations. Using the R command cor(), we get low correlation ($r = 0.152$) between the two measurements (Figure 14).

4.3 Intraclass Correlation Coefficient

To explore the variability of the predicted peak velocities at time A and time B, we calculate the Intraclass Correlation Coefficient (R_{icc}) using the definition by Fleiss [5]. According to Fleiss, our study is "a simple replication reliability study conforming to a one-way random effects model." An estimator for R_{icc} in this case is:

$$\hat{R}_{icc} = \frac{s_T^2}{s_T^2 + s_e^2},$$

where s_T^2 is an unbiased estimator of σ_T^2, the variance component "due to error-free variability among subjects," and s_e^2 is an unbiased estimator for σ_e^2, "the estimated component of variance due to the random measurement error." The two components of variance are calculated using the following formulas:

$$s_T^2 = \frac{\text{BMS - WMS}}{k}$$

$$s_e^2 = \text{WMS},$$

where BMS is the Between-subject Mean Square and WMS is the Within-subject Mean Square.

The following table is borrowed from Fleiss [5] to illustrate the relationships among variables for the analysis of variance in our study:

	degrees of freedom	**sum of squares**	**mean square**
Between patients	$N-1$	$\sum_{i=1}^{14} k(\bar{X}_i - \bar{X})^2$	BMS
Within patients	$K-N$	$\sum_{i=1}^{14}(k-1)s_i^2$	WMS

where N is the number of subjects, k is the number of measurements per person, K is the total number of measurements for all participants ($K = k \cdot N$), $\bar{X}_i$ is the mean for participant i, $\bar{X}$ is the mean for all participants ($\bar{X} = \sum_{i=1}^{14} \bar{X}_i/N$), s_i is the variance for participant i, and BMS and WMS are the corresponding sum of squares divided by degrees of freedom:

$$\text{BMS} = \frac{\sum_{i=1}^{14}(\bar{X}_i - \bar{X})^2}{N-1}$$

$$\text{WMS} = \frac{\sum_{i=1}^{14}(k-1)s_i^2}{K-N}.$$

Our numerical results are:

	degrees of freedom	**sum of squares**	**mean square**
Between patients	13	2490.69	191.59
Within patients	14	1937.80	138.41

Then $s_T^2 = (191.59 - 138.41)/2 = 26.59, s_e^2 = 138.41$, and $\hat{R}_{icc} = 26.59/(26.59 + 138.41) = 0.16$. This implies that 16% of the variance between the predicted peak velocities at amplitude 1° in times A and time B results from inter-individual difference. That is, 84% of the variance between the two variables are due to intra-individual variability and errors.

The approximate one-sided $100(1-\alpha)\%$ confidence interval for R_{icc} [5] is

$$R_{icc} \geq \frac{\frac{\text{BMS}}{\text{WMS}} - F_{N-1,K-N,\alpha}}{\frac{\text{BMS}}{\text{WMS}} + (k_0 - 1)F_{N-1,K-N,\alpha}},$$

where $F_{v_1,v_2,\alpha}$ is the tabulated value of the F distribution with v_1 and v_2 degrees of freedom. In our study, $F_{14-1,2\cdot 14-14,0.05} = F_{13,14,0.05} = 2.51$. Hence,

$$R_{icc} \geq \frac{\frac{191.59}{138.41} - 2.51}{\frac{191.59}{138.41} + (2-1)\cdot 2.51} = -0.289.$$

The confidence interval of R_{icc} includes 0, which, according to Fleiss [5] means that we accept the hypothesis that the underlying value of R_{icc} is 0. That is, the differences between subjects are "due exclusively to random measurement errors." The low confidence interval bound for R_{icc} also indicates poorer reliability of the data that the R_{icc} implies.

4.4 Number of Participants Needed

When a test is carried out only once, the minimum number of participants required in each group to detect differences between the experimental group and the control group is given by Fleiss [5] as:

$$n_1 = \frac{2(\sigma_T^2 + \sigma_e^2)(z_{\alpha/2} + z_\beta)^2}{\delta^2}, \tag{4}$$

where σ's are variances estimated by s's as in the last section, δ is the desired difference in mean of the two groups, and $z_{\alpha/2}$ and z_β are the $\alpha/2$ and β fractiles of the standard normal distribution. In general, $\alpha = 0.05$ and $\beta = 0.05$, and correspondingly, $z_{\alpha/2} = 1.96$ and $z_\beta = 1.65$. $z_{\alpha/2}$ is our control of type I error, the probability of detecting a significant result when none exists and z_β is our control of the type II error, the probability not detecting a significant result that is actual present. With the estimate of the variance and tolerance levels for alpha and beta, we can estimate the sample size n required to see a difference of δ if a difference exists.

An important way to increase the reliability of test results is to conduct replicate experiments on the same sample. As suggested by Fleiss [5], if we use $\bar{X}_m$ to represent the mean of m replicate measurements, then $Var(\bar{X}_m) = \sigma_T^2 + \sigma_e^2/m$. Hence the intraclass correlation coefficient with m replications is

$$R_m = \frac{\sigma_t^2}{\sigma_T^2 + \sigma_e^2/m} = \frac{m \cdot R}{1 + (m-1)R},$$

where R is the intraclass correlation coefficient when there is no replication in the experiment. Observe from this formula when $m > 1$, $R_m > R$, and as m increases, R_m becomes closer to 1. This implies that as the number of replication tests increases, intraclass correlation coefficient increases, and the share of inter-individual variance in the total variance increases as well.

The minimum sample size for each group required to detect a significant difference is

$$n_m = \frac{2(\sigma_T^2 + \frac{\sigma_e^2}{m})(z_{\alpha/2} + z_\beta)^2}{\delta^2} \tag{5}$$

It could be observed from (5) that n_1 is always greater than n_m, which aligns with the hypothesis that, compared to a single test, replication of a test decreases the minimum sample size required for each group to detect a difference.

A computationally more convenient formula to use is:

$$n_m = \frac{n_1 R}{R_m} = \frac{n_1(1 + (m-1)R)}{m}.$$

In our example, using 53.177 and 276.828 as estimates for σ_T^2 and σ_e^2 respectively, $n_1 = 344$. This implies that 344 participants are needed for each group in order to detect a significant difference between patients and controls. Increasing the number of replications to 2, $n_2 = 200$. So 200 participants are required for each group. When replicating the test three times, $n_3 = 152$, and n_m will continue to decrease as m increases.

5 Results for other variables

We apply the analysis in the previous section to other variables of the sham group and summarize all results along with results from the previous section in the table below:

variable	**Shapiro**[2] time A time B	**t**[2]	**Wilcoxon**[2]	**r**	$\hat{R}_{icc}$	**confidence interval (≥)**	n_1	n_2	n_3
peak velocity									
asymptote	0.134 0.0234	-	0.562	0.525	0.530	0.129	1086	830	745
predicted value at 1°	0.680 0.839	0.417	0.398	0.152	0.161	-0.289	173	100	76
predicted value at 5°	0.920 0.0107	-	0.112	0.151	0.166	-0.284	88	52	39
peak acceleration									
asymptote[3]	0.662 0.0633	0.523	0.711	0.685	0.682	0.340	135	114	107
predicted value at 1°	0.756 0.0179	-	0.0873	0.0436	0.0716	-0.370	177	95	68
predicted value at 5°	0.345 0.0242	-	0.0366	0.282	0.294	-0.156	112	73	60
duration									
asymptote[4]	0.446 0.430	0.815	0.936	0.414	0.426	-0.00515	9	7	6
predicted value at 5°	0.664 0.0356	-	0.787	0.409	0.371	-0.0707	61	42	36

Note: [1]p-values are reported. [2]One participant's result at time B is an outlier and so is taken away from the tests. [3]Asymptote and predicted value at 1° are the same given the model for duration and amplitude.

Data for asymptotes of the function for peak velocity and amplitude are from fittings for individuals discussed earlier. Variables describing relationships between peak acceleration vs. amplitude and duration vs. amplitude are adopted from previous work [4]. The variables are generated using the same models that we need (function (1) for peak acceleration vs. amplitude and function (2) for duration vs. amplitude) using least squares fit in MATLAB.

6 Discussion

There is wide variation in the literature on methods of analyzing eye-tracking data. For example, Leigh and Zee [7] combine all saccades from multiple individuals to develop an exponential model for analysis, and Cifu et al. [4] fit exponential models to each participant. Bollen et al. [2], who also fit models of the main sequence relationships for each individual, use different functions, the power function, for fitting. This study explores the variability of saccadic relationships (following similar methods as Bollen et al. [2]) with the data in Cifu et al. [4]. In particular we focus on the variables asymptotes and predicted 1° and 5° for peak velocity, peak acceleration, and duration.

As shown in the table in the results section, in general, intraclass correlation coefficients are higher for asymptotes than predicted values at 1° and 5°, meaning that asymptotes reflect higher inter-individual variances and smaller intra-individual variances than predicted values at 1° and 5°.

The lower bounds for confidence intervals of R_{icc}'s for asymptotes are higher than those for predicted values at 1° and 5° as well, supporting the idea that asymptotes are more reliable variables than the other two. Confidence intervals for intraclass correlation coefficients of predicted value at 1° and 5° for peak velocity, peak acceleration, and duration are all negative, meaning that we are unable to reject the hypothesis that the underlying value of R_{icc} is 0 and that the differences between test results of the two times are due exclusively to random errors.

On the other hand, the minimum group sizes that allow detection of differences between test results at the two times are smaller when using predicted values at 5° than using predicted values at 1° and asymptotes for peak velocity and peak acceleration. It implies that when we only have a limited number of participants in our experiment, it is easier to detect test differences if we analyze the hypothetical values of peak velocity and peak acceleration at 5°. On the other hand, for the relationship between duration and amplitude, if we use asymptote or the hypothetical duration at 1°, we need fewer participants than if we use the hypothetical duration at 5°. This may be due to the fact that we use the exponential function (1) to model the relationships between peak velocity vs. amplitude and peak acceleration vs. amplitude, and we use the power function (2) for duration vs. amplitude.

Whichever variable we pick and whichever relationship we examine, replication of tests remarkably decreases the minimum number of participants needed in each group to detect test differences. When group size is small, conducting multiple replicate tests may be another way to reduce measurement error. Replication helps to reduce intra-individual variance, so if there is a difference between the two sets of variables, we are more likely to find it. If no difference exists, given not finding a difference from repletion of tests, we are more certain that there is no difference. Although replication of tests also incurs additional costs and efforts, comparing with the usually even higher costs of gathering a larger number of participants for a single test, replication of tests is still the more economic approach. Therefore, we recommend replicating the same test at least once when designing new eye-tracking experiments.

References

[1] Bahill AT, Clark MR, Stark L. The Main Sequence, A Tool for Studying Human Eye Movements. *Mathematical Biosciences* 24, 191-204 (1975) American Elsevier Publishing Company, Inc.

[2] Bollen E, Bax, J, van Dijk JG, Koning M, Bos JE, Kramer CGS, van der Velde EA. Variability of the Main Sequence. *Investigative Ophthalmology & Visual Science.* 1993, Vol.34, No.13. pp 3700-3704.

[3] Brenner LA, Terrio H, Homaifar BY, Gutierrez PM, Staves PJ, Harwood JE, Reeves D, Adler LE, Ivins BJ, Helmick K, Warden DF. Neuropsychological test performance in soldiers with blast-related TBI. *Neuropsychol* 2010;24(2):160-7.

[4] Cifu DX, Wares JR, Hoke KW, Wetzel PA, Gitchel G, Carne W, *Differential Eye Movements in Mild Traumatic Brain Injury vs. Normal Controls*

[5] Fleiss JL, *The Design and Analysis of Clinical Experiments* New York: John Wiley and Sons, 1986.

[6] Guskiewicz KM, Ross SE, Marshall SW. Postural Stability and Neuropsychological Deficits After Concussion in Collegiate Athletes. J Athl Train 2001;36(3):263-273.

[7] Leigh RJ, Zee DS. *The Neurology of Eye Movements.* Oxford University Press, Oxford, UK, fourth edition, 2006.

[8] Murkin JM, Arango M, Near-infrared spectroscopy as an index of brain and tissue oxygenation. *Br J Anaesth* 2009; 103 (suppl 1): i3-i13.

[9] Nesselroade JR, Ram N. Studying Intraindividual Variability: What We Have Learned That Will Help Us Understand Lives in Context. *Research In Human Development* 1(1&2), 9-29.

[10] Pickett TC, Radfar-Baublitz LS, McDonald SD, Walker WC, Cifu DX, Objectively assessing balance deficits after TBI: Role of computerized posturography. *J Rehabil Res Dev* 2007; 44(7): 983-90.

[11] The parameters of eye movement. `http://www.liv.ac.uk/~pcknox/teaching/Eymovs/params.htm`