Statistical Methods in Research

Microsurgery Study

Final Project Report

May 4, 2018

Compiled By:

Hosein Neeli (1541673)

Bharat Verma (1639951)

Pushpendra (1639812)

Guided By:

Prof. Dr. Ioannis Pavlidis

George Panagopoulos (TA)

University of Houston, Main Campus

Spring Semester, 2018

Contribution

	Module.Name	Executed.By
1	Introduction	Bharat
2	Normalization	Pushpendra
3	Task Vs Time Analysis	Hosein
4	Perinasal Perspiration	Bharat
5	Surgeon Comparison	Bharat
6	Task Score Analysis	Pushpendra
7	Nasa-TLX	Hosein
8	LinearRegression	Pushpendra and Hosein
9	Project Report Writing	Bharat, Hosein, Pushpendra
10	Project Report Review	Pushpendra
11	Presentation	Pushpendra and Hosein
12	Data Gathering using R	Pushpendra
13	Graph modelling	Bharat and Hosein
14	Quality Control questions 1,2 and 3	Pushpendra
15	Quality Control question 4	Bharat
16	Quality Control question 5	Hosein

Its now been long that Statistical Analysis have been used to provide solutions to complex problems in such diverse areas as communications, stability, finding patterns in data sets of variables and other areas of interest. Although these problems previously lacked adequate mathematical treatment, the results of statistical analysis have been significant both for explanation of a pattern and also for prediction. Statistical analysis in general is a scientific method and it has applications in many areas of scientific research.

This project consists of implementing a statistical analysis using R, which provides patterns to explain stress and performance relationship during the microsurgical task of cutting and suturing in an inanimate simulator by medical students. The project uses statistical analysis methods like linear regression, correlation, hypothesis and AOV(analysis of variance), non-parametric test(Wilkinsons) test to analyze the relationship between the stress and level of performance.

The main feature of the project is input data set from Methodist hospital and as Statistical student using R Programming language we performed all the analyses and test. We have also generated graphical representation and tabular summarization of the analysis results.

1 Introduction

We studied 22 medical students who participated in a longitudinal study regarding the relationship of sympathetic stress arousal and skill in learning micro-surgical tasks. One way to evaluate stress is using perinasal perspiration measurement. This method has been applied by Methodist hospital to assess effect of stress on performance of Microsurgery students in two delicate tasks of cutting and suturing. In this project, we analyze given data of Methodist hospital and describe the relevance of stress and several other parameters to performance of the students.

For this experiment each subject had to pay five visits, lasting one hour each, in order to practice micro-surgical cutting and suturing in an inanimate simulator. In their first visit, and after signing an informed consent, the subjects completed a NASA-TLX questionnaire, and a trait anxiety inventory(TIA) form. At the end of their last visit they completed a post-study questionnaire.

1.1 Objectives

In this study, we have analyzed the effect of stress to the performance of 15 out of 22 medical surgery students in two separate tasks of cutting and suturing. We considered only those 15 candidates for whom we received all the required performance related information. In addition we analyzed the improvement in performance, amount of stress of the students during the five session of surgical training. We are presenting all our analysis visually using graph.

Besides this we are also providing graphical information for all the 22 students in Appendix section of this report.

1.2 Protocol Standards

A protocol is the precise and detailed design for conducting a research study, approved IRB Protocol was used to do this study. An IRB is an appropriately constituted group that has been formally designated to review and monitor biomedical research involving human subjects. In accordance with FDA regulations, an IRB has the authority to approve, require modifications in (to secure approval), or disapprove research. This group review serves an important role in the protection of the rights and welfare of human research subjects.

The purpose of IRB review is to assure, both in advance and by periodic review, that appropriate steps are taken to protect the rights and welfare of humans participating as subjects in the research. To accomplish this purpose, IRBs use a group process to review research protocols and related materials (e.g., informed consent documents and investigator brochures) to ensure protection of the rights and welfare of human subjects of research.

Thin films are used to simulate human skin for performing cutting and suturing tasks.

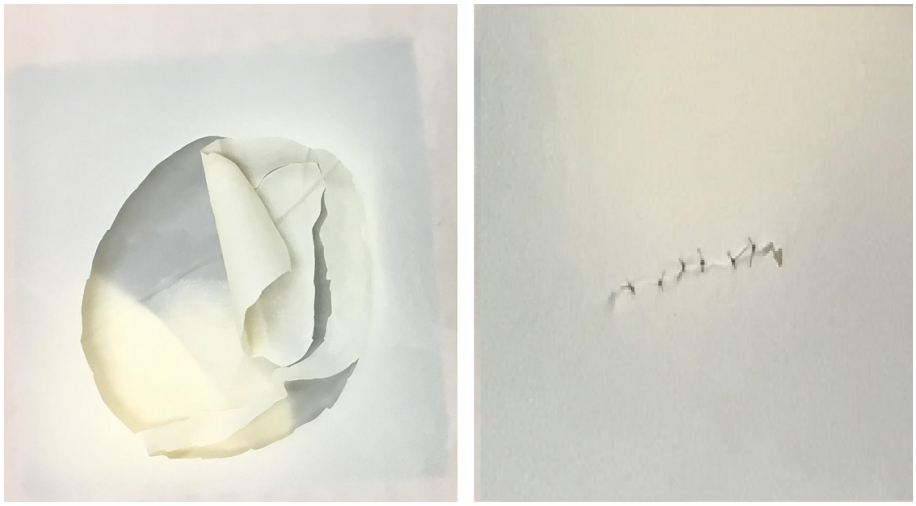

Figure 1: Left: sample of cutting. Right: sample of suturing.

1.3 Study Layout and Task

The experiment starts by first task of measuring the perinasal perspiration (stress) during relaxing for five minutes, where subjects listen to spa music. Next they perform two dexterous task of cutting and suturing and measurement is done for perinasal perspiration during each task here as well.

Listed below are the tasks which should be performed by each subject in each of the five sessions:

- Baseline task: Subjects relax and listen to 5 minutes of relaxing music in order to reach a constant level of calmness.
- Cutting task: Subjects had to precision cutting in the inanimate simulator. They were facially recorded by a thermal and visual camera.
- Cutting questionnaire: Subjects had to fill out a questionnaire called NASA-TLX for cutting task. This questionnaire instrument features five sub scales measuring different aspects of the subjects' perceptions regarding task difficulty.
- Suturing task: The subjects had to perform suturing in the inanimate simulator. They were facially recorded by a thermal and visual camera.
- Suturing questionnaire: Subjects had to fill out a questionnaire called NASA-TLX for suturing task. This questionnaire instrument features five sub scales measuring different aspects of the subjects' perceptions regarding task difficulty.

1.4 Population of Study

The study was done on 15 medical students between 22 and 26 (average age is 23.1 years), about 73% of them are newbie and just completed 1 year of Microsurgical Specialty experience. Among these student, 5 are females and the rest 10 are males.

Note: This whole study took around 9 months of period to complete so there is a change in the age of some of the students when the study started and by the time it got completed. The id of such students are 1, 4, 22 and 24.

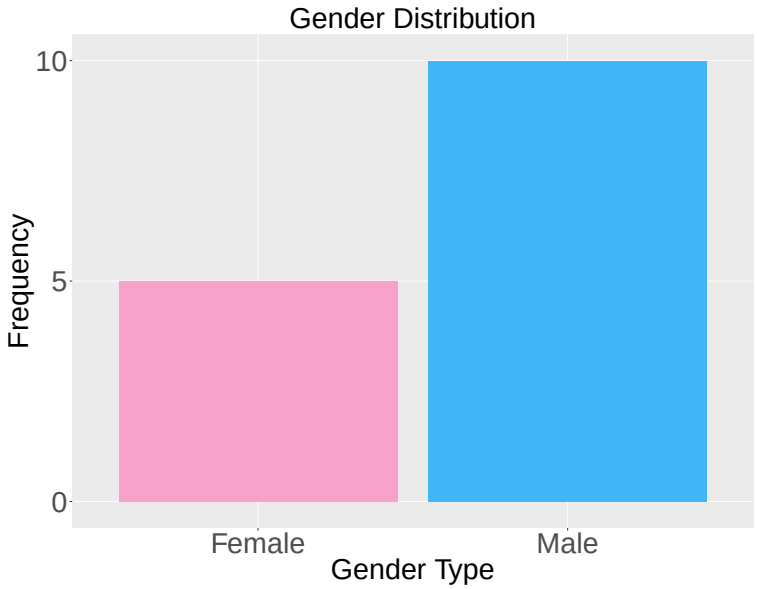

Figure 2: Histogram of Gender Distribution.

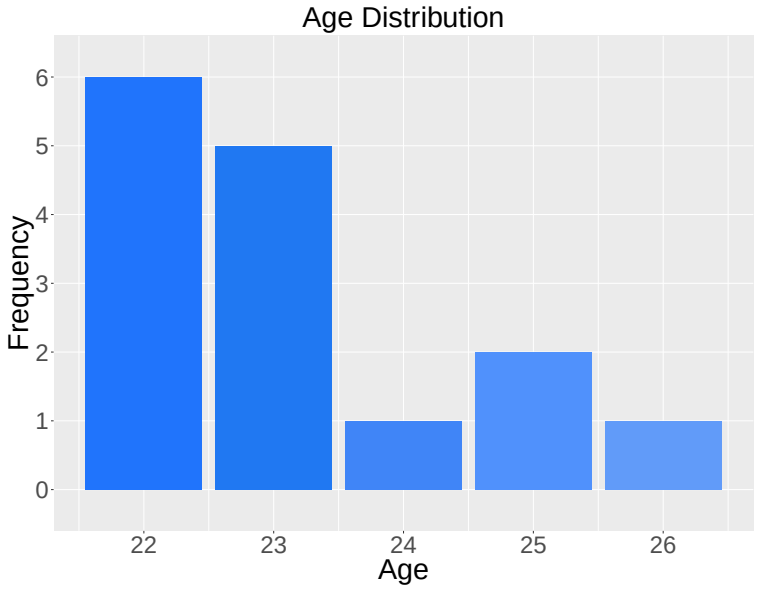

Figure 3: Histogram of Age Distribution.

1.5 Duration of Study

This study was conducted during five sessions over a period of around 9 months. Each session consist of 1 hour and subject is supposed to perform their duties and answer to the questionnaires.

1.6 Trait Psychometric Data (TIA)

Trait Psychometric Data (TIA) takes values in the range 20-80, with scores up to low 40s considered normal, while higher scores considered indicative of overanxious individuals.

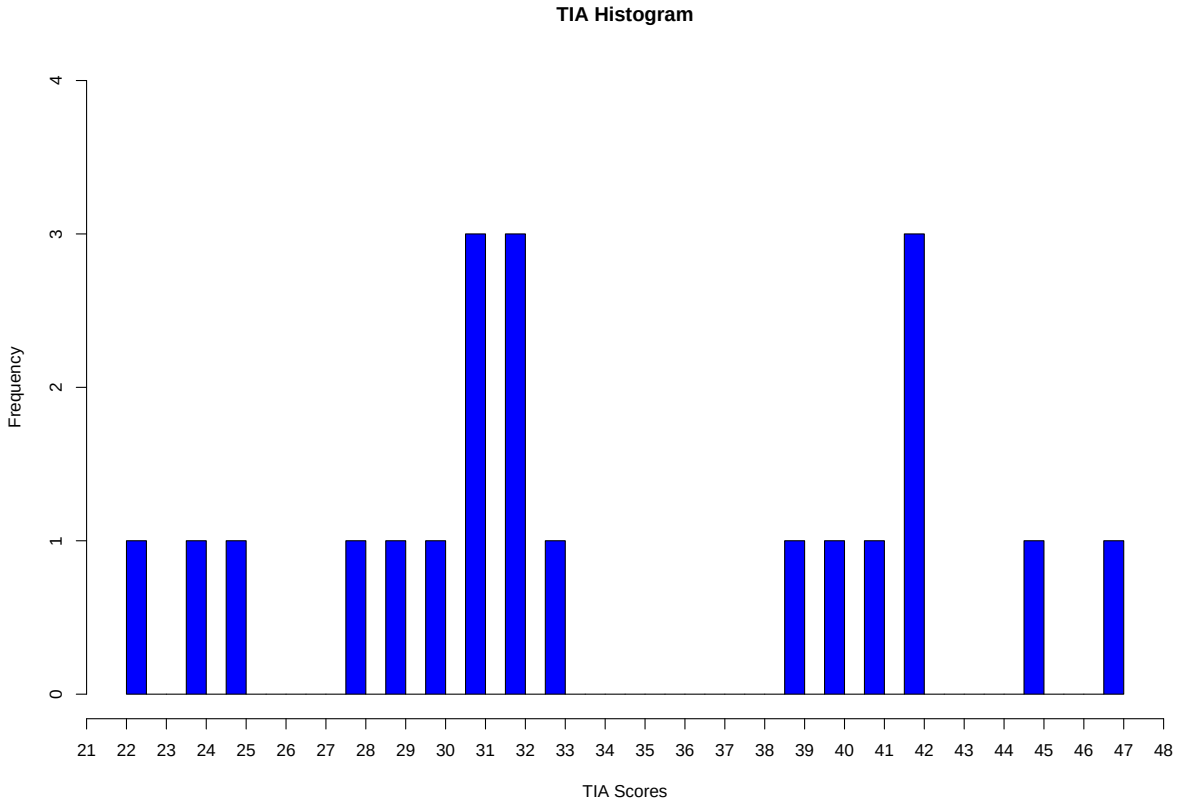

Figure 4: Histogram of TIA Scores.

1.7 Perinasal Perspiration Measurement

Perinasal perspiration is the amount of sweat around nasal area of a human. Perinasal perspiration is increased when a human doing a delicate and hard task undergoes pressure or stress, so we use the amount of change in perinasal perspiration to measure the stress level a person undergoes during the conduct of experiment.

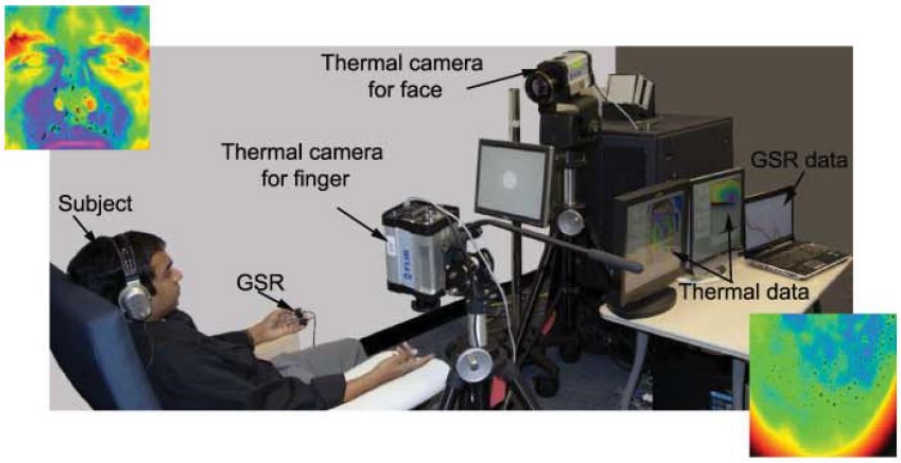

Figure 5: Perinasal perspiration recording process during baseline (relaxing) task. Two thermal cameras measure perinasal and index finger thermal changes of the subject.

A sensitive thermal device records the amount of perinasal perspiration of the subject during baseline (relaxing), cutting and suturing tasks couple of time in a second from the start to the end of each session. These data have been used to evaluate the stress level of each subject and to determine which task is more difficult for subjects.

The experiment also employed thermal sensor device which measures perinasal perspiration of the subjects and record the values during both the micro surgical task.

1.8 Microsurgery Performance Sessions

Microsurgery Performance Sessions consist of two task: Cutting and Suturing.

Cutting: In cutting task, subjects had to cut a thin inanimate simulator with precision.

Suturing: In suturing task, they had to sutur the already cutted inanimate simulator a maximum of six times in a row and maximum time alloted for this is 20 minutes.

Figure 6: Left, sample of cutting. Right, sample of suturing. Thin films are used to simulate human skin for performing cutting and suturing tasks

The next goal is the evaluation of work done by the students. For this two surgeons separately scored the performance of these tasks based on their judgment. The details of their methodology to grade the student's tasks is not known to us.

All the measurements and scores were given to the Data Expert team to work on it.

1.9 Student Questionnaires

The Student Questionnaire instrument features six sub scales measuring different aspects of the subjects' perceptions regarding task difficulty. These sub scales are:

- Effort
- Mental Demand
- Physical Demand
- Frustration
- Performance
- Temporal Demand

2 Processing Data and Analysis

2.1 Significance level

Before we run any statistical test, it is important to first define or set the alpha level, which is also called the "significance level". By definition, the alpha level is the probability of rejecting the null hypothesis when the null hypothesis is true. As it is medical field so in this study of analysis, we consider the α value as 0.01 in all test.

2.2 Normality Test of Data

This section we are going to test for the normality of data.

Cutting time: By looking at the cutting time we can visually see that it is varying a lot among subjects and among sessions, compare to suturing times that almost are 20:00 minutes.

Suturing time: We can see that the suturing time is most of the time among subjects and among sessions remains to 20:00 minutes.

To make sure that the difference between values come from a normal distribution, we are going to perform normality test for both cutting time as well as suturing time.

We assume the generalized hypothesis as:

- H_0: The data comes from a normally distributed source.
- H_1: The data doesn't come from a normally distributed source.

Results from R:

Table 1: Shapiro-Wilk normality test for cutting time.

Session No.	p-Value	Null-Hypothesis	p-Value Normalized
Session 1	0.2492	Accepted	0.3077
Session 2	0.1252	Accepted	0.5049
Session 3	0.1454	Accepted	0.3346
Session 4	0.1088	Accepted	0.7593
Session 5	0.005595	Rejected	0.1167

From the result we get that just 5th session data is not normal in actual data but gets normalized after applying transformation using squareroot.

Lets look at the quantile plot

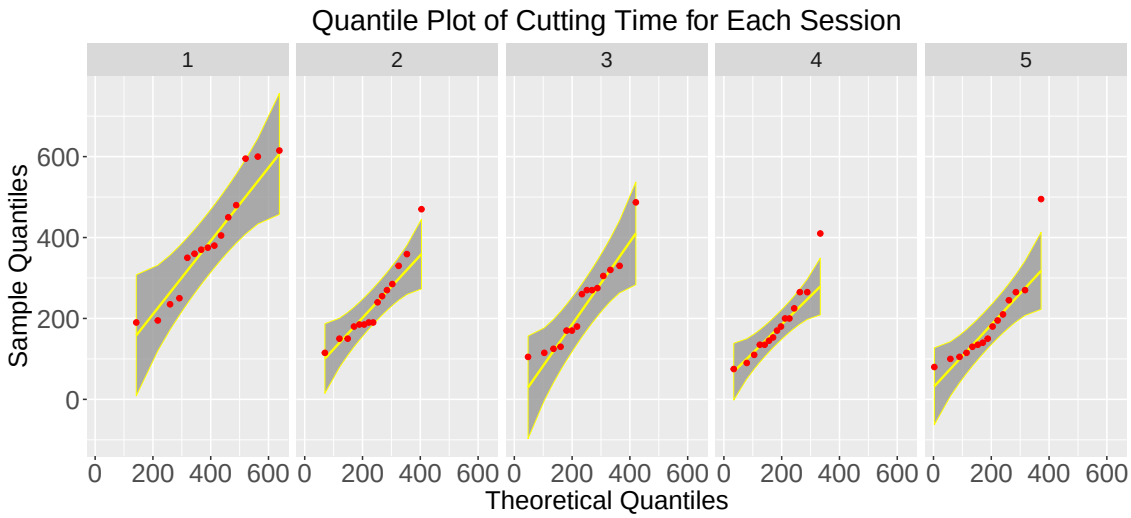

Figure 7: Quantile plot of cutting time per sessions

Table 2: Shapiro-Wilk normality test for suturing time.

Session No.	p-Value	Null-Hypothesis	p-Value Normalized
Session 1	9.834e-08	Rejected	9.834e-08
Session 2	1.046e-05	Rejected	1.014e-05
Session 3	7.102e-05	Rejected	6.543e-05
Session 4	0.0009085	Rejected	0.00080
Session 5	0.008363	Rejected	0.003681

From the result we get that the suturing time is not all normal in any of the sessions. So to normalize it we used squareroot and this makes it better but still not normal.

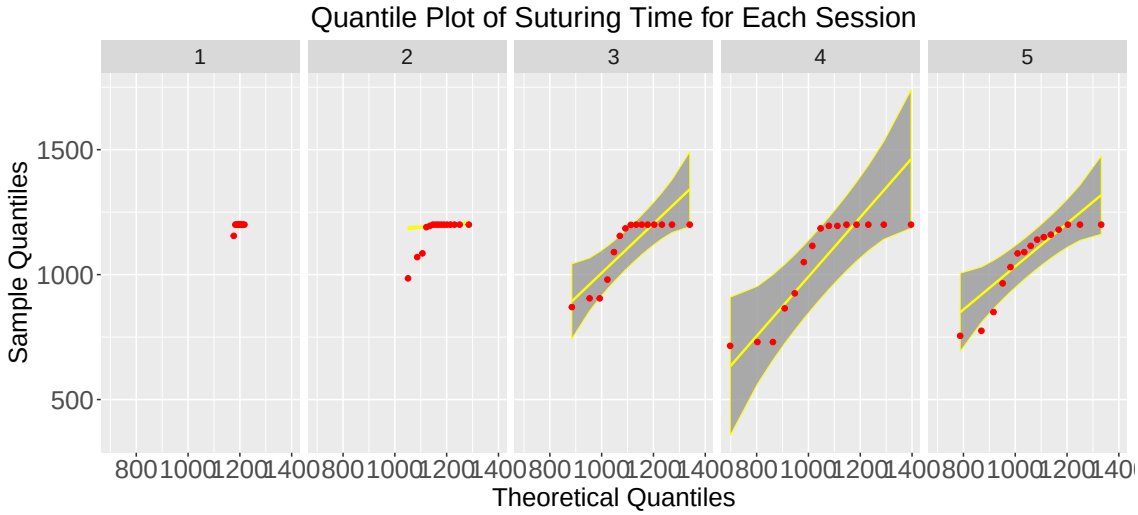

Figure 8: Quantile plot of suturing time per sessions

We are going to evaluate normality of data for perinasal perspiration, the hypothesis we have defined is:

- H_0: The data comes from a normally distributed source.
- H_1: The data doesn't come from a normally distributed source.

Here are the results:

Table 3: Shapiro-Wilk normality test for difference of cutting and suturing by baseline task perspiration values.

	Session	Cutting	Null-Hypothesis	Suturing	Null-Hypothesis
1	Session 1	0.003734	Rejected	0.004928	Rejected
2	Session 2	0.01295	Accepted	0.02522	Accepted
3	Session 3	0.2396	Accepted	9.29e-09	Rejected
4	Session 4	0.1349	Accepted	0.04248	Accepted
5	Session 5	1.22e-08	Rejected	3.77e-05	Rejected

Normalization of perspiration for linear regression model 1) Subtract the Cutting/Suturing perspiration from Baseline

2) We get some negative values in part(1), to remove all negative numbers we find the minimum negative number in (1) and add it to every element in this set.

3) Due to adding the minimum number we get some zero values, to make it working with log we added 0.001.

2.3 Task versus Time Analysis

We are now analyzing the improvement in time as part of the learning process during the five session for cutting and suturing for all subjects. We define a generalized hypothesis here for both the task cutting and suturing:

Given below is the aggregated plot of cutting and suturing time for all subjects in five session experiment.

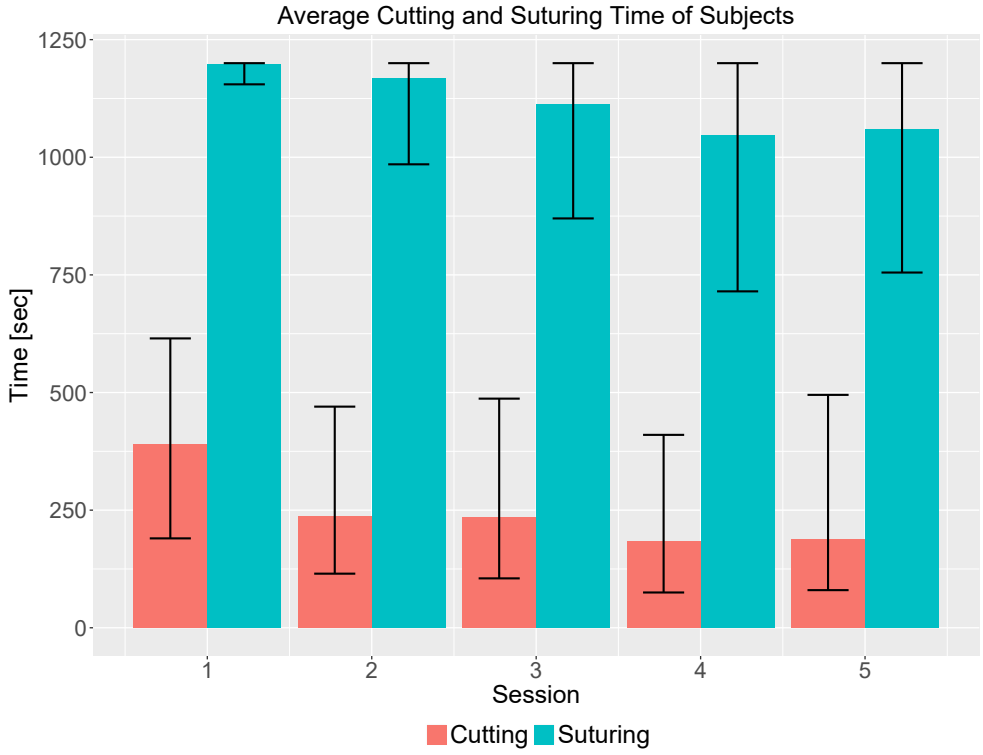

Figure 9: Aggregated plot of cutting and suturing time of all subjects. Whiskers indicate range of each task in each session.

- H_0: There is no significant change (increase or decrease) in average task time of the subjects in five experimental sessions (in other words average task time of every session is equal).
- H_1: There is significant change in average task time for the subjects in at least two (or more) sessions (in other words the average task time of at least two (or more) session are significantly different).

We did Kruskal-Wallis test for both cutting and suturing time to analyze the if there is any difference between task time of the subjects in each session.

Table 4: Kruskal-Wallis test result for cutting and suturing for performance time.

Task	p-Value	Null-Hypothesis
Cutting	0.0001739	Rejected
Suturing	0.00047	Rejected

Result and analysis:

The p-values from of Kruskal-Wallis test indicates that null hypothesis for both cutting and suturing tasks are rejected, so we can infer here that mean cutting and suturing performance time has improved as the experiments progressed. So, the learning process successfully enhanced subject's speed in both task.

Furthermore, the average cutting time for subjects decreased from 390 seconds in first session to 188 seconds in the last session. It shows that their performance in cutting task improved by 51%.

In case of suturing task, the average suturing time for subjects decreased from 1197 seconds in first session to 1060 seconds in the last session. Here the improvement is only about 10%. It shows that the suturing is a harder task than cutting.

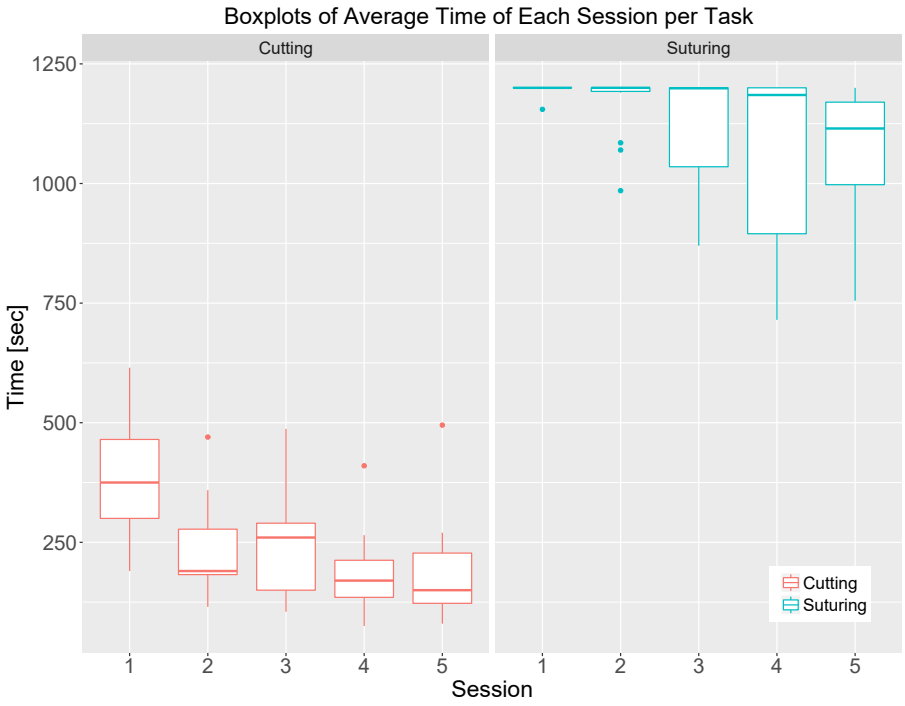

Figure 10: Box plot of average time of sessions, categorized by cutting and suturing task.

As we can see in the figure[19], almost none of the subject could complete their task in the first two sessions, but from 3rd session, we can see the improvement in suturing task started. For cutting task,

improvement in performance of the subjects started immediately from first session, it shows that the cutting session is relatively easier for them to perform.

Important Note: It does not make sense to use suturing time as is for analyzing the performance of the subjects. The reason is by looking at the data, we can see that subjects in almost all of the sessions have average time performance near 20:00 minutes. We can not infer anything special with just including all 20:00 minutes value in our model and inferences. So we came up with a new term called "Suture speed" or "Suture Quantity". As we have the number of successful suture in each session by each subject, we can easily calculate the time a subject needs to complete 1 successful suture by this equation:

$$SutureSpeed = \frac{TotalSuturingTime}{NumberofSutures}$$

Table 5: Kruskal-Wallis test result for cutting and suturing for performance time. (Corrected suturing time)

Task	p-Value	Null-Hypothesis
Cutting	0.0001739	Rejected
Suturing	0.0004114	Rejected

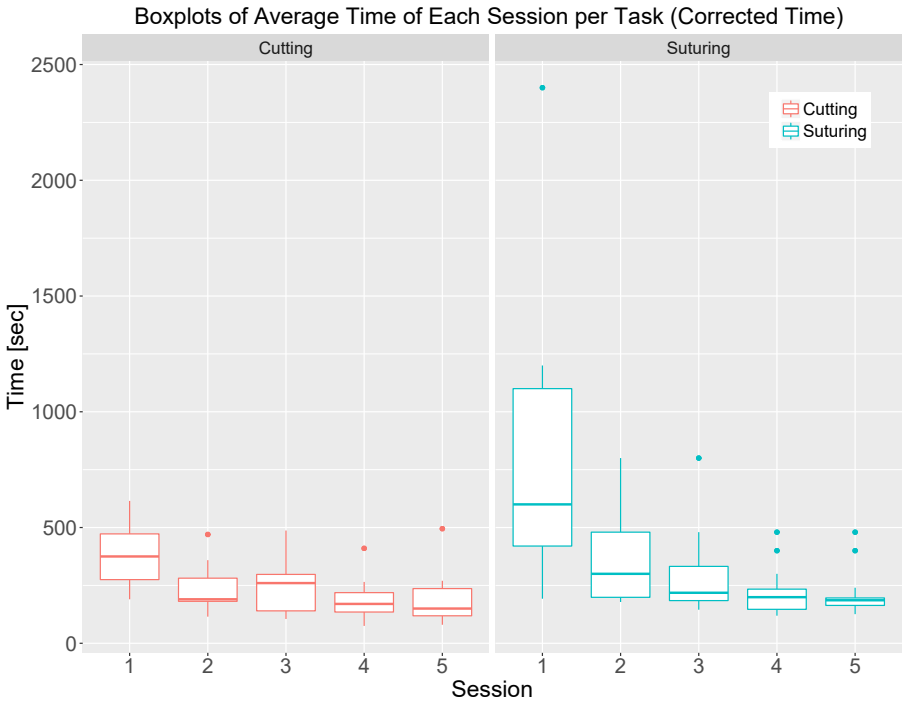

Figure 11: Box plot of average time of sessions, categorized by cutting and suturing task. The suturing time indicates the required time for 1 successful suture

Another reason that this new term is more useful than the traditional one is that by looking at scores of the students we can see that the score of suturing task does not depend on time it takes to complete a suturing session, it more depends on the number of successful sutures. So for analyzing time it's more reasonable to use time of each suture instead of overall time.

Looking at the difference between boxplots and the pattern of decreasing time has changed significantly. Additionally we observe there are some exceptions among the subjects:

- Specifically for suturing, You can see that the range of times of suturing in the first session were very small, however, this ranges increases significantly until the last session. The reason that maximum

time for suturing remained the same is that there are some subjects whose suturing time hadn't improved during five sessions and remained on 20:00 minutes. As an example, in figure[19] we can see that suturing time of subjects 1, 12 and 13 remained almost 20:00 minutes for all sessions.

In contrast, we can see that the minimum time of suturing decreased from 1st to last session. It shows that some subjects have improvement in the time of suturing, however, some other subjects do their suturing task significantly faster compared to previous sessions. The minimum Suturing time among all sessions is 11 minutes and 50 seconds which belongs to subject 9 in 4th session.

- By looking at the aggregated plot at figure [19], we can see that subject No. 24 had deteriorated and the time of cutting and suturing for this subject increased from first to last session. We can not infer that this subject has problems in learning abilities, the reason of the bad performance might be personal issues during the period of experiment. We recommend that special attention is required for this subject.
- Subject No. 24 in the 4th session had the best performance in cutting task with 1 minute and 30 seconds. In contrast, subject No. 3 in the firs session had the worst time performance in cutting task with 10 minutes and 15 seconds.
- Best timing performance in suturing task belongs to subject No. 9 in the 8th session with 11 minutes and 55 seconds
- From the aggregated plot, we can see that subjects No. 1, 2, 12, 13 and 22 had almost no progress in suturing task. Further investigation is needed to find out what was the problem that caused them not to do very well among all sessions of experiment.

2.4 Perinasal Perspiration Analysis

Figure 12: Box plot of baseline, cutting and suturing perinasal perspiration (stress) for each session.

- Figure 12 display's the overall trend in the perinasal perspiration for each session.

  For baseline we can see that for first session perinasal perspiration is higher compare to all other session, the reason for the same could have been because they were just introduce to this process

so the perinasal perspiration was higher, from the next session they became more comfortable so the perspirations also became stable in each session.

- For cutting we can see the upward trend in the over all perspiration during each session, initially subject might not be aware of the difficulty of the cutting task once they realize that perinasal perspiration increases.
- For suturing we can see upward trend in session 3 and session 4 ,we are not sure what causes the increase in perinasal perspiration for these two session.

2.5 Surgeon Comparison

The task done by students are graded by two surgeons based on their defined parameters. We are interested here to find out if method of scoring of two surgeons are the same. If it is not, we will face problem with reliability of data. The reason is that the score is the main parameter that we can use to evaluate the performance of the students and so if the mean of the scores given by the surgeons are different, we can not trust the effect of other parameters to the scores.

To make sure that two surgeons have same method of scoring, we are going to perform paired T-Test for each session of cutting and suturing.

We form the generalized hypothesis as:

- H_0: The means of the scores graded by two scorers are not significantly different. $(\mu_0 = \mu_1)$
- H_1: The means of the scores graded by two scorers are significantly different. $(\mu_0 \neq \mu_1)$

Result obtained from paired T test in R:

Table 6: T-Test for scores by each scorer

Session No.	Task	p-Value	Null-Hypothesis
Session 1	Cutting	1	Accepted
	Suturing	0.1306	Accepted
Session 2	Cutting	0.2654	Accepted
	Suturing	0.1271	Accepted
Session 3	Cutting	0.6102	Accepted
	Suturing	0.1232	Accepted
Session 4	Cutting	0.8178	Accepted
	Suturing	0.3704	Accepted
Session 5	Cutting	0.5264	Accepted
	Suturing	0.8116	Accepted

From the above result we get that all of the hypothesis in each task and for each session are accepted. It means in all of the sessions the means of the scores given by each surgeon are not significantly different and so we are confident that they have a similar method of scoring. If scoring methods of surgeons are different, we would have to dig further to find out the reasons that their performance seem not equal from the scorers perspective.

2.6 Task Score Analysis

In section we analyze accuracy score of the subjects in cutting and suturing.

We define the generalized hypothesis as:

- H_0: There is no significant change between the average score of the subjects during the experiment session. $(\mu_{ScoreOfSession1} = \mu_{ScoreOfSession2} = \mu_{ScoreOfSession3} = \mu_{ScoreOfSession4} = \mu_{ScoreOfSession5})$

- H_1: There is significant change of average score of the subjects at least between two (or more) experiment session. ($\mu_{ScoreOfSession1} \neq \mu_{ScoreOfSession2} \neq \mu_{ScoreOfSession3} \neq \mu_{ScoreOfSession4} \neq \mu_{ScoreOfSession5}$)

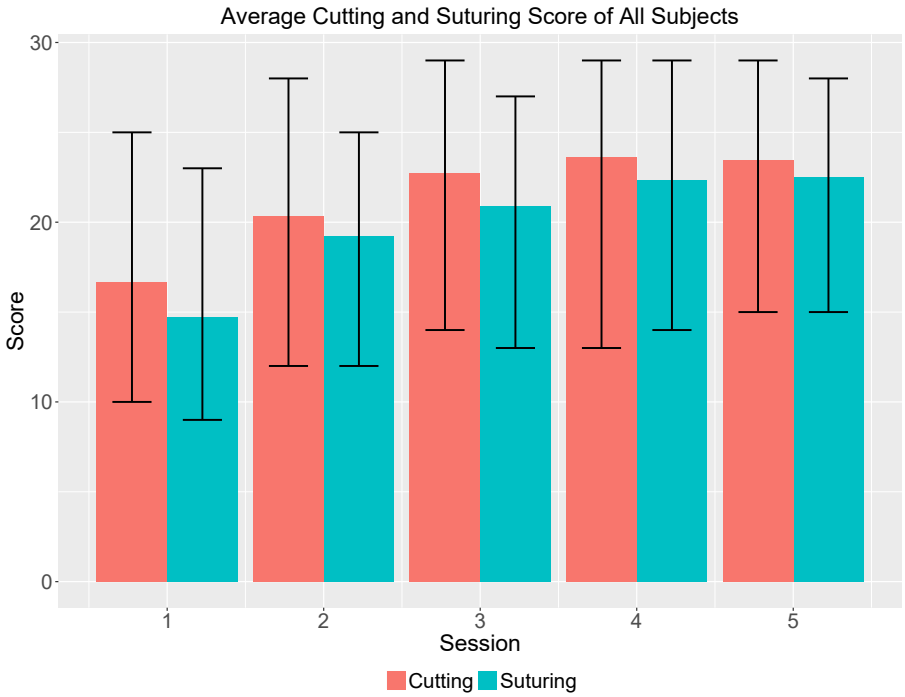

Figure 13: Average cutting and suturing score of all subjects. Whiskers indicates range score for each task in each session.

From the above aggregated score plot of the subjects for both cutting and suturing we observe that overall they have progressed. The average of cutting score in each session is higher than suturing score. It shows that cutting task is relatively easier.

Additionally looking into the average score of 4th and 5th session for both the task we find that it is same. From we can infer the below:

- Their performance in cutting and suturing task converged to a consistent level.
- They were no more interested in the experiment as it spanned over longer period of time.
- The surgeons who graded the students might not be satisfied with the students performance and so they have implemented tougher ways to evaluate the students work.

Table 7: Kruskal-Wallis test result for cutting and suturing accuracy score.

Task	p-Value	Null-Hypothesis
Cutting	9.137e-10	Rejected
Suturing	1.993e-10	Rejected

By looking at p-values of the Kruskal-Wallis test of accuracy score for tasks for all sessions, we can see that both cutting and suturing accuracy improved significantly from beginning to the end of the experiment. This is the proof for effectiveness of process of microsurgery drill sessions.

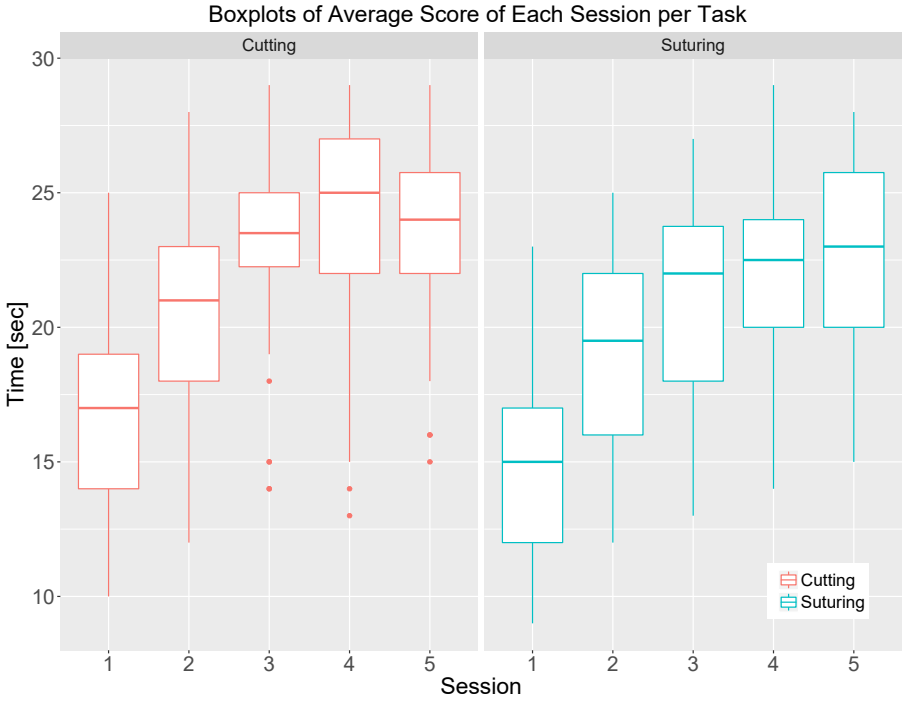

Figure 14: Box plot of average accuracy score of sessions, categorized by cutting and suturing task.

Let's look at the plot fig[25] of score to go deep into the details.

- Subject 24 has a noticeably loss the score from 4th to 5th session. From task time data we inferred that this subject might have problems during the test period maybe because of some external factors.
- Subject 7 from the 1st session had very consistent performance and got high scores.
- Subjects 22 and 4 had improvements in the score in every session.

2.7 Nasa-TLX Questionnaires Results Analysis

This data is collected as part of feedback from all subjects. We are going to analyze from this information if there is any significant progress made during sessions. To analyze this data we use non-parametric Kruskal-Wallis Test. This test is used to find whether there is a significant difference between sample population mean of more than two populations.

We form the generalized hypothesis for all six responses as:

- H_0: The performance, physical demand, temporal demand, mental demand, effort and frustration (each separately) is constant among all the sessions and there are no change in these parameters (Progress or regress). $(\mu_1 = \mu_2 = \mu_3 = \mu_4 = \mu_5)$
- H_1: The performance, physical demand, temporal demand, mental demand, effort and frustration (each separately) decrease or increase in at least one of the sessions $(\mu_0 \neq \mu_1 or \mu_0 \neq \mu_2 or \mu_1 \neq \mu_3$ ...and so on.

Results from R:

Table 8: Kruskal-Wallis test result for NASA-TLX feedback questionnaire

Response.	Task	p-Value	Null-Hypothesis
Mental Demand	Cutting	0.006955	Rejected
	Suturing	0.1841	Accepted
Physical Demand	Cutting	0.03913	Accepted
	Suturing	0.07722	Accepted
Temporal Demand	Cutting	0.4368	Accepted
	Suturing	0.7659	Accepted
Performance	Cutting	0.4314	Accepted
	Suturing	0.0006136	Rejected
Effort	Cutting	0.002083	Rejected
	Suturing	0.3499	Accepted
Frustration	Cutting	0.001269	Rejected
	Suturing	0.05207	Accepted

Box plot for cutting Nasa-TLX responses:

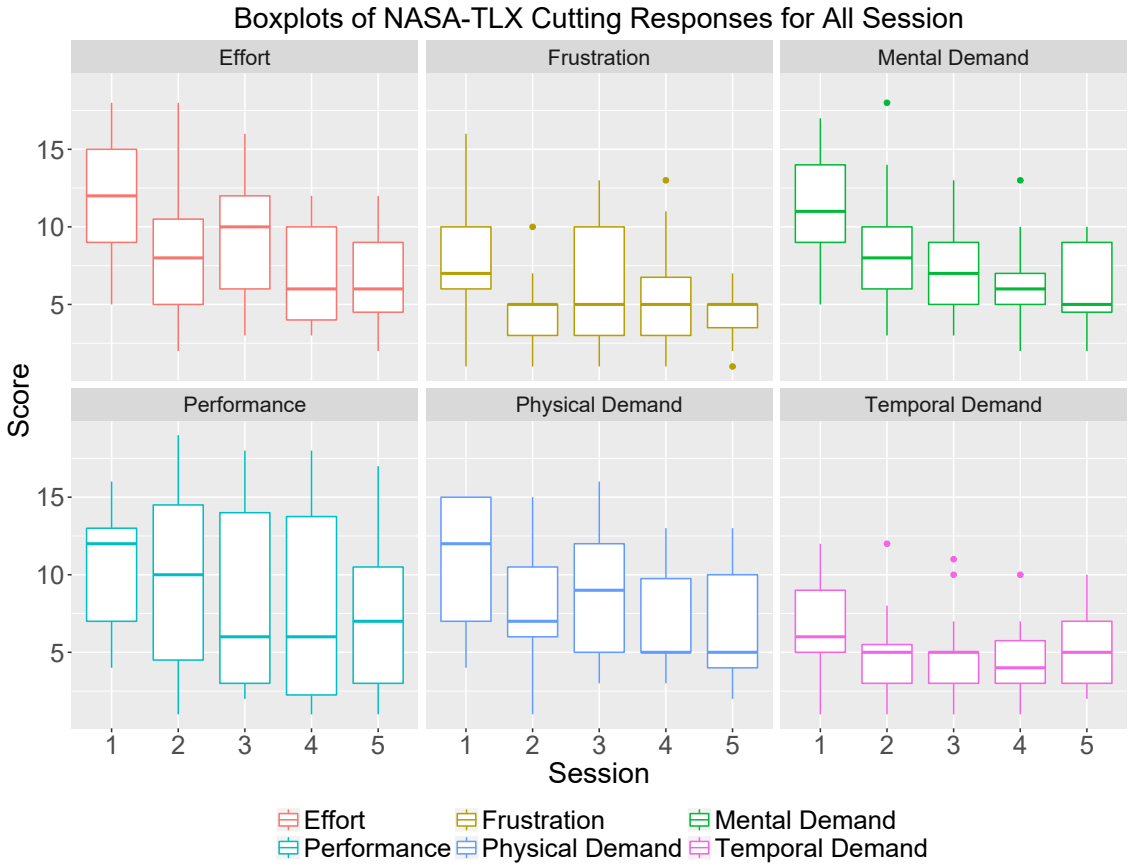

Figure 15: Distribution of mean score (points) per session of cutting task. Mean values of Effort, mental demand and frustration are significant between first to fifth session.

Box plot for suturing Nasa-TLX responses::

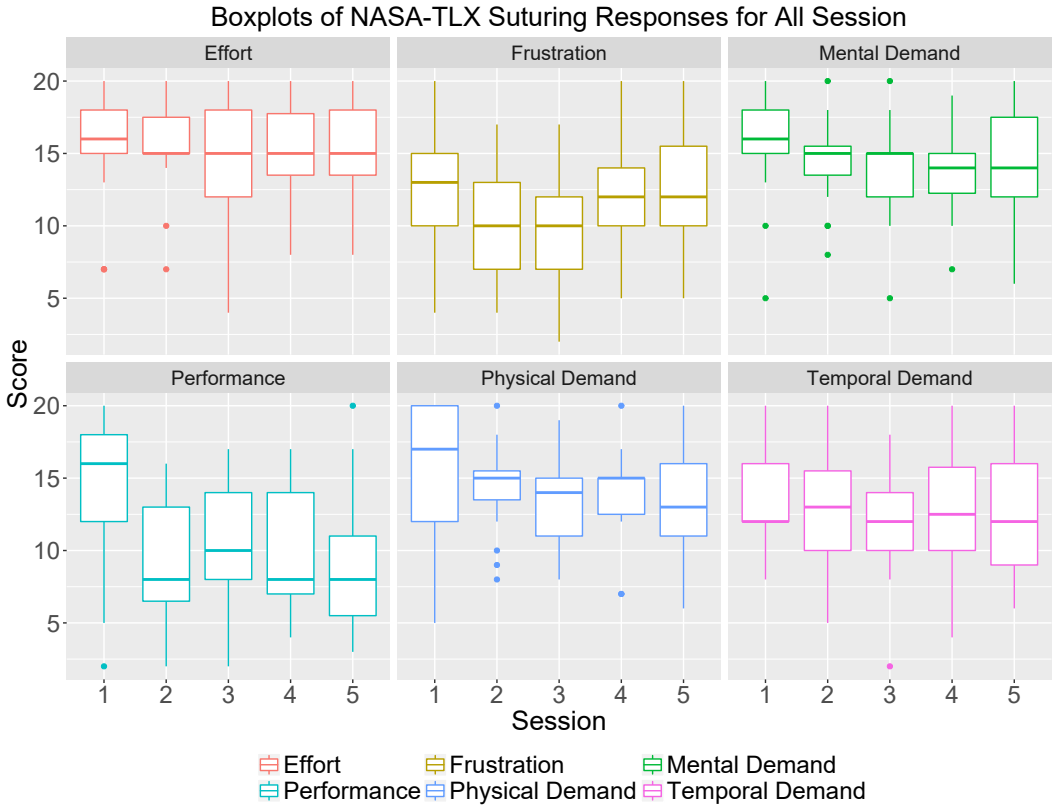

Figure 16: Distribution of mean score (points) per session of suturing task. Mean score of performance is significant between first to fifth session.

As we can see the R result in the table above for p-values (assumed $\alpha = 0.01$), the null hypothesis are accepted in all categories except Cutting Mental Demand, Suturing Performance and Cutting Frustration. It means that subjects believe that there was a change between mental demand from the first to the last session, and as we can see in the plot, this change is downward so they believe the task of cutting in latter sessions needs less mental demand. Same justification can be provided for Cutting effort and suturing Frustration, they think these tasks become easier in latter sessions and need less effort and cause less frustration.

One of the reason why the mean score of questionnaire for suturing performance has a decreasing pattern might be because the subjects in the first session may have assumed that the suturing task is not very challenging and they can do it in pretty decent way, but after the first experience, they realized their performance and balance their expectations.

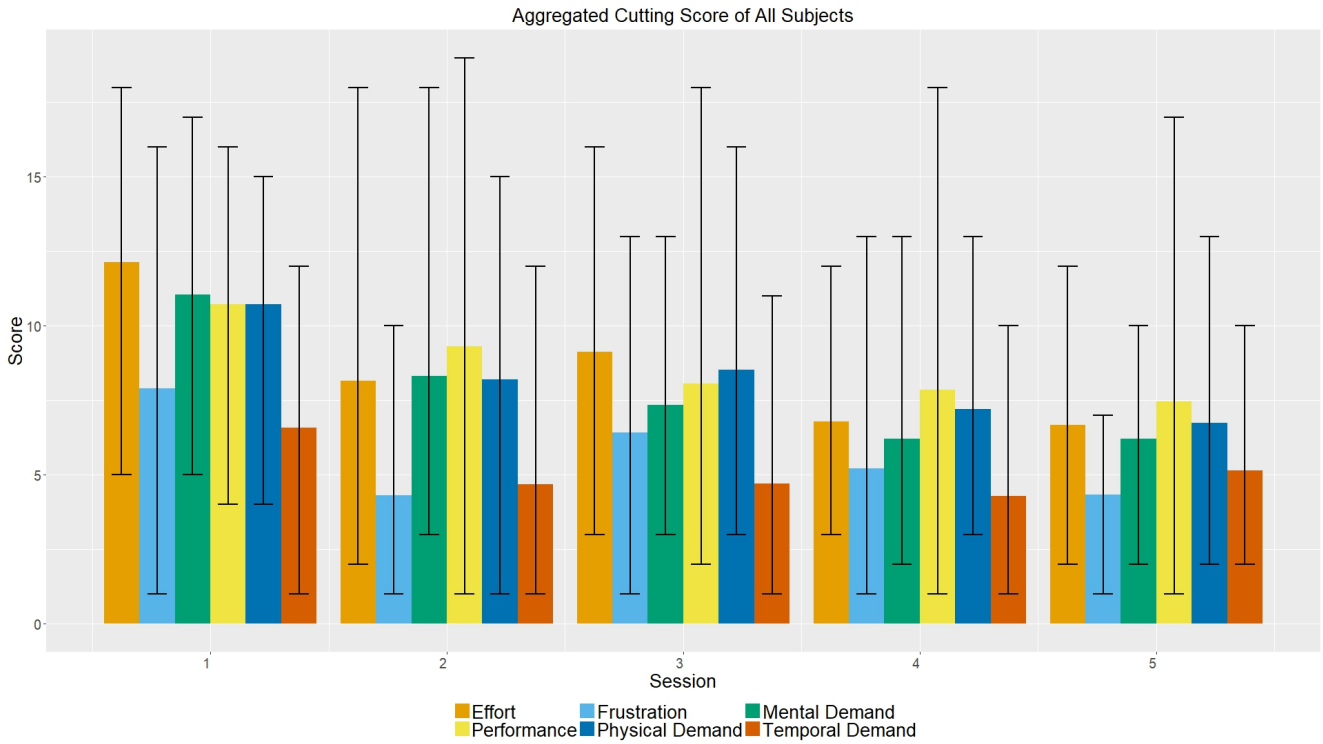

Figure 17: Aggregated plot of average cutting score of NASA-TLX questionnaire of all subjects. Whiskers indicates range between minimum and maximum value in each section.

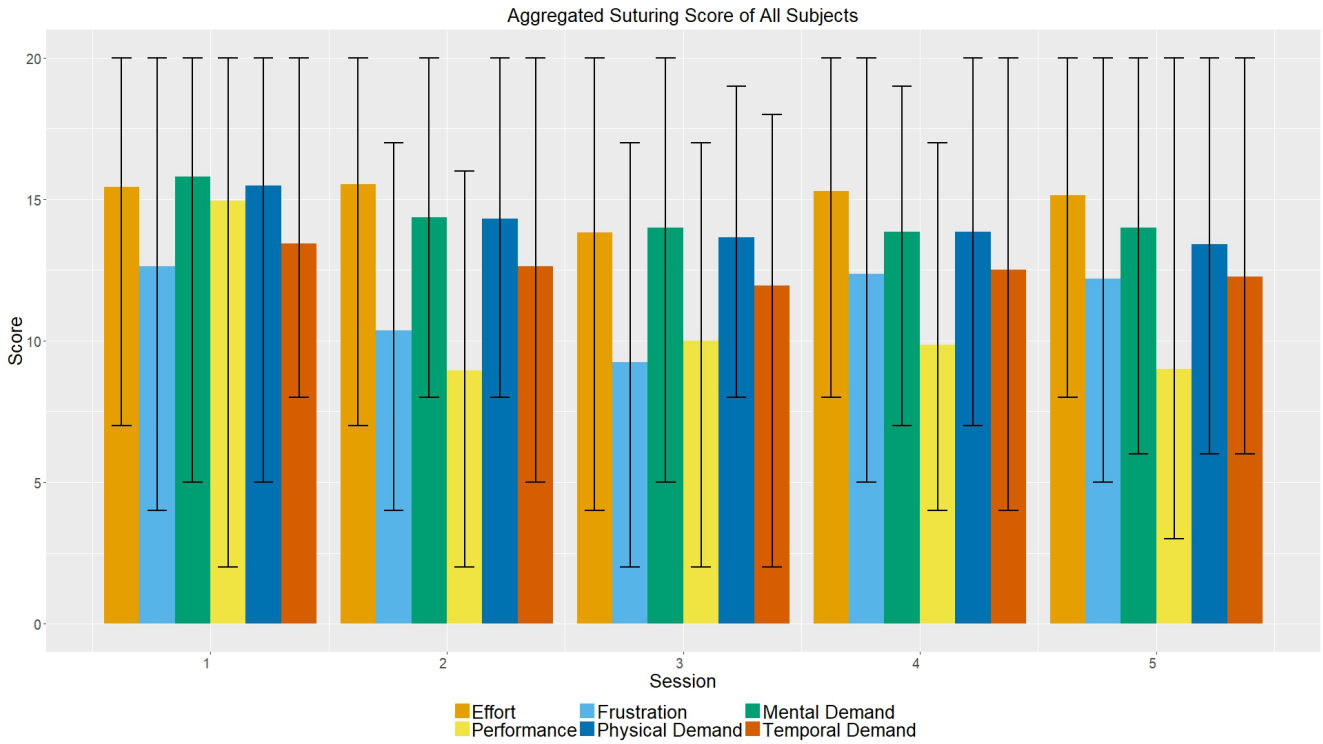

Figure 18: Aggregated plot of average suturing score of NASA-TLX questionnaire of all subjects. Whiskers indicates range between minimum and maximum value in each section.

As we can see in Figure 22. and 23. for subject No. 4 we observe that during the whole study period a descending pattern in physical, mental pressure, frustration and effort. The important point is that we can see downward pattern in "performance" scores. It shows that maybe the procedure of training is not very effective for this specific user or may be there some problem with the subject.

3 Statistical Inference and Models

In this section, we are going to inference data using statistical models. The main concept of this session is based on linear regression.

3.1 Estimation of score by other parameters

To estimate the performance of the subjects, we use accuracy score. By applying a linear regression model we can estimate this accuracy score based on some independent variables to predict score of a subject, to see whether whether performs well or not.

We form the linear regression for estimating score:

$$LM(Score \tilde{} Task + Session + Perspiration + Scorer + Age + Sex + Time + Session * Sex)$$

The hypothesis testing is:

- H_0: There is no significant effect on performance accuracy score by session, stress (perinasal perspiration), scorer, age, sex, time of task and interaction between any of them.

- H_1: There is at least one parameter among session, stress (perinasal perspiration), scorer, age, sex, time of task and interaction between any of them that has significant effect on performance score.

Result from R:

Table 9: Estimates of linear model for accuracy Score

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	8.4630	5.0443	1.68	0.0946	
TaskSuturing	-1.1125	0.7131	-1.56	0.1199	
Session2	5.7385	1.1940	4.81	2.59e-06	***
Session3	6.6864	1.1970	5.59	5.79e-08	***
Session4	7.6297	1.2148	6.28	1.39e-09	***
Session5	7.7551	1.2163	6.38	8.12e-10	***
Scorer2	0.0072	0.4123	0.02	0.9861	
Age	0.4371	0.1678	2.60	0.0097	**
SexMale	1.1297	1.0581	1.07	0.2866	
Normalized PPlog	-0.4502	0.5940	-0.76	0.4491	
Normalized Time	-0.2312	0.0550	-4.20	3.59e-05	***
Session2:SexMale	-4.2311	1.4097	-3.00	0.0029	**
Session3:SexMale	-3.0753	1.4043	-2.19	0.0294	*
Session4:SexMale	-3.1965	1.4202	-2.25	0.0252	*
Session5:SexMale	-2.9130	1.4274	-2.04	0.0423	*
Observations	140				
R^2	0.4947				
Adjusted R^2	0.4678				
Residual Std. Error	3.437 (df = 263) 263				
F Statistic	18.39*** (df = 8; 286)				
p-value:	2.2e-16				

The overall p-value =2.2e-16 shows that the accuracy score significantly affect by at least one of session, stress (perinasal perspiration), scorer surgeon, age or sex parameters.

In this model, the score of the subjects is the response variable (dependent variable) which is estimated by task, session, scorer, age, sex and time as independent variables plus interaction between session and sex. We did tested for interactions between other parameters and found them to be insignificant and they didn't affect the other coefficients significantly. Here is the analysis of the result of this model:

- The task in this model has no significant effect on the accuracy score, the reason is that we used a new defined parameter as suture quality. Session of microsurgery has a very strong significant effect on the accuracy score of the subject, as we can see in the linear regression model result, there is a positive increasing effect of session. It means that in each session, the accuracy score of the subject is increased significantly. We can see that if we make the first session as the baseline (No added point considered for session 1 to the result), in session 2 the subject will get about 5.7 points more than session 1, in session 3 compared to session 1, subject will get 6.68 more points (0.98 points more compared to the last session). And 7.62 points for session 4 and 7.75 for session 5. The jump in accuracy score from session 2 to 3 and 3 to 4 is higher than the increase of score from session 4 to 5, it means that the performance in session 4 slightly the same as the performance in the last session.
- There will be no significant effect in accuracy score if the scorer 1 or scorer 2 grade the subject.
- The age makes significant difference in accuracy score, and this effect is positive. We can infer that the older subject, or we can call them more experienced subjects get significantly more accuracy score compared to younger ones. The difference of score is about 0.43 score for increasing each year in age.
- The task performance time plays an important role in accuracy of the subjects. The negative coefficient of the time shows that each seconds of delay in completing the task decreases 0.23

points from the accuracy score of the subject and this amount makes a significant effect on the overall accuracy score.

- The perinasal perspiration (stress) of the subjects has no significant effect on their performance. The p-value for perinasal perspiration is considerably high and we reject the null hypothesis that there is any significant effect of stress on accuracy performance of the subjects in cutting and suturing.
- By looking at the coefficient of the interaction of gender and session, we can infer that overall performance of the male subject are lower than female subjects in each session. A male subject gets 4.23 lower score in the first session than a female subject (considering every other parameter is constant and same), this phenomenon is repeated for the other sessions, but the difference is slightly balanced for the latter sessions as we can check the difference of the coefficients.
- We can conclude that the sessions, the age of the subject, the time they spend for the task and the interaction of gender and session make significant changes in accuracy score of the subject. This score will not be significantly affected by type of task, the scorer or gender of the subject.

3.2 Estimation of Stress by other Parameters

To estimate the amount of stress, we define a linear regression model to evaluate effect of task, session, age and sex on the perinasal perspiration. The reason to do this estimation is to find out which independent parameter has significant effect on stress during microsurgery. This result will help surgery teachers to balance the parameters in a way that cause smaller amount of stress for a student of surgery. However, in previous section we saw that the accuracy score was not affected by stress, but from personal perspective, stress can harm the health of a person so the less stress helps to keep the surgeon in better state of health so they can perform their job for longer period of time.

We form the linear regression as:

$$LM(Stress(PP)\tilde{}Session + Task + Age + Sex)$$

Result from R:

Table 10: Estimates of linear model for Cutting Score

	Estimate	Std. Error	t value	Pr(>|t|)	
(Intercept)	-4.7000	0.4141	-11.35	2e-16	***
TaskSuturing	-0.9604	0.0427	-22.52	2e-16	***
Session2	0.0305	0.0686	0.45	0.6566	
Session3	0.1906	0.0674	2.83	0.0050	**
Session4	0.1818	0.0685	2.65	0.0084	**
Session5	0.0557	0.0692	0.81	0.4214	
Age	0.0273	0.0173	1.58	0.1147	
SexMale	-0.0917	0.0466	-1.97	0.0500	*
Observations	140				
R^2	0.4947				
Adjusted R^2	0.4678				
Residual Std. Error	0.3555 (df = 270) 263				
F Statistic	75.42^{***} (df = 8; 270)				
p-value:	2.e-16				

- From the coefficients of the model we can see that the task has very significant effect on stress. The suturing task has more effect on stress rather than cutting task.
- In addition, Session 3 and 4 have slight effect on stress. These two sessions should be reviewed to find out what was the cause of stress in those two sessions.

- We can see that age has not any significant effect on stress, but gender has. Male subjects tend to be more stressful that female subjects in this case.

4 Discussion

We recommend the following:

- The order of task in experiment must be randomized. This will help us analyze if there is any impact of cutting on suturing or reverse.
- The proportion of males and females should be statistically equal. Using this we can better find out if there is any impact of sex on the experiment.
- More information must be revealed to us in order to form and infer our linear regression in a better way.
- The given dataset has missing data at many places, if it can be avoided we could more confident and consistent in our analysis.

5 Experimental Design

This was an experimental study in which 22 medical students who participated in a longitudinal study regarding the relationship of sympathetic stress arousal and skill in learning micro-surgical tasks. A protocol is the precise and detailed design for conducting a research study, approved IRB Protocol was used to do this study. This study was conducted during five sessions over a period of around 9 months. Each session consist of 1 hour and subject is supposed to perform their duties and answer to the questionnaires.

The experiment starts by first task of measuring the perinasal perspiration (stress) during relaxing for five minutes, where subjects listen to spa music. Next they perform two dexterous task of cutting and suturing and measurement is done for perinasal perspiration during each task here as well.

Listed below are the tasks which should be performed by each subject in each of the five sessions:

- Baseline task: Subjects relax and listen to 5 minutes of relaxing music in order to reach a constant level of calmness.
- Cutting task: Subjects had to precision cutting in the inanimate simulator. They were facially recorded by a thermal and visual camera.
- Cutting questionnaire: Subjects had to fill out a questionnaire called NASA-TLX for cutting task. This questionnaire instrument features five sub scales measuring different aspects of the subjects' perceptions regarding task difficulty.
- Suturing task: The subjects had to perform suturing in the inanimate simulator. They were facially recorded by a thermal and visual camera.
- Suturing questionnaire: Subjects had to fill out a questionnaire called NASA-TLX for suturing task. This questionnaire instrument features five sub scales measuring different aspects of the subjects' perceptions regarding task difficulty.

Perinasal perspiration is increased when a human doing a delicate and hard task undergoes pressure or stress, so we use the amount of change in perinasal perspiration to measure the stress level a person undergoes during the conduct of experiment. A sensitive thermal device records the amount of perinasal perspiration of the subject during baseline (relaxing), cutting and suturing tasks couple of time in a second from the start to the end of each session. These data have been used to evaluate the stress level of each subject and to determine which task is more difficult for subjects. The experiment also employed thermal sensor device which measures perinasal perspiration of the subjects and record the values during both the micro surgical task.

6 Appendix

6.1 Plots

This section is to show Quality Controlled plots. In this section, we represent data that contain detailed information about various parameters of the subjects in each session.

Figure 19: Cutting and suturing time of all subjects.

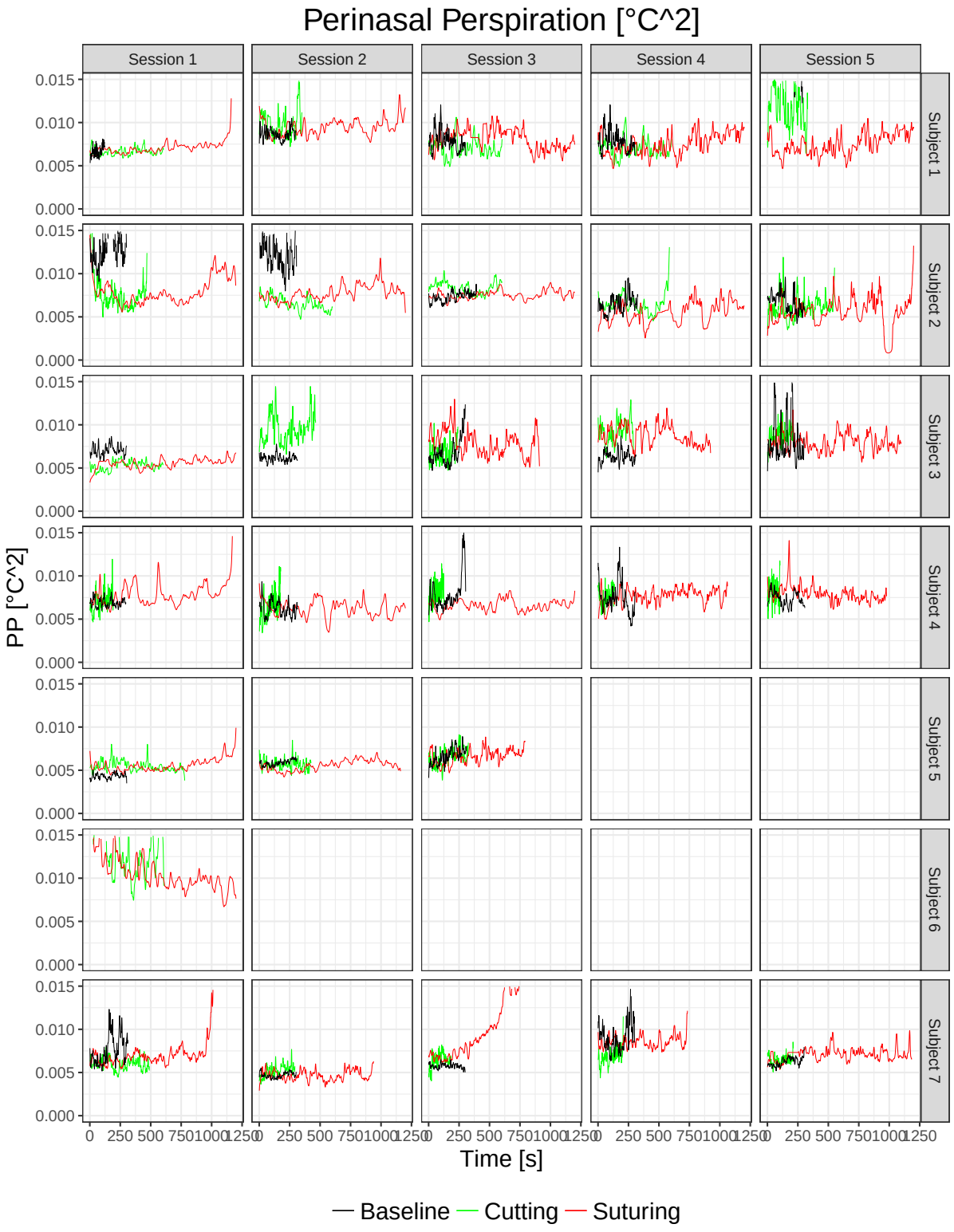

Figure 20: Perinasal perspiration of subjects in each session. For some sessions, there are missing data because of malfunction of thermal camera or other reasons.

Figure 21: Perinasal perspiration of subjects in each session. For some sessions, there are missing data because of malfunction of thermal camera or other reasons.

Figure 22: Perinasal perspiration of subjects in each session. For some sessions, there are missing data because of malfunction of thermal camera or other reasons.

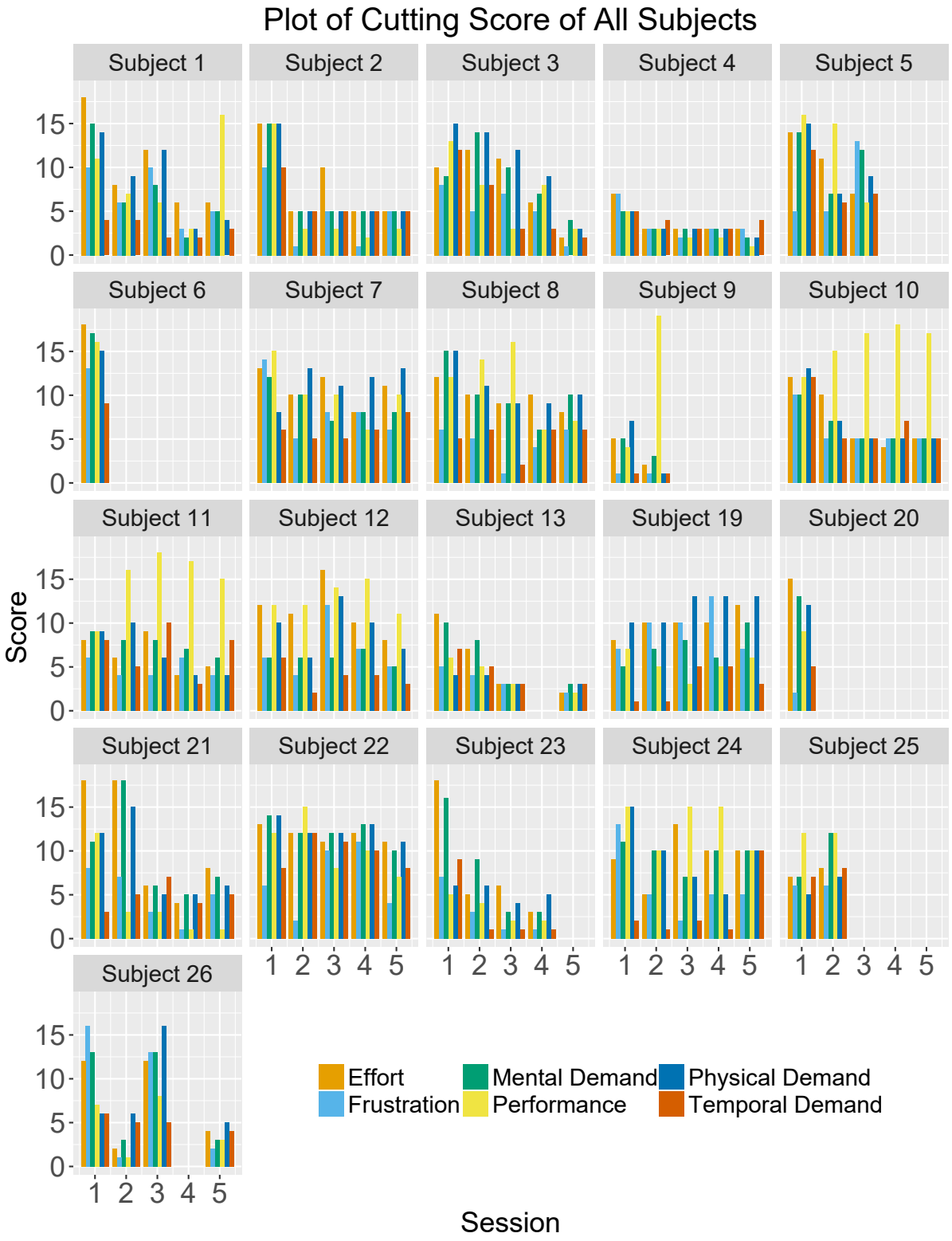

Figure 23: Aggregated plot of cutting score of all subjects.

Figure 24: Aggregated plot of suturing score of all subjects.

Figure 25: Cutting and suturing score of all subjects.

6.2 Data Adequacy Information

We provide here the missing data information in the given dataset to us session wise for each subject. The 0 in figure indicates missing data while 1 means data is present.

NASA-TLX Data

Table 11: Data availability of NASA-TLX questionnaire, 1 indicates that data is available, 0 indicates data is not available.

	Subject	Session 1	Session 2	Session 3	Session 4	Session5
1	Subject1	1	1	1	1	1
2	Subject2	1	1	1	1	1
3	Subject3	1	1	1	1	1
4	Subject4	1	1	1	1	1
5	Subject5	1	1	1	0	0
6	Subject6	1	0	0	0	0
7	Subject7	1	1	1	1	1
8	Subject8	1	1	1	1	1
9	Subject9	1	1	0	0	0
10	Subject10	1	1	1	1	1
11	Subject11	1	1	1	1	1
12	Subject12	1	1	1	1	1
13	Subject13	1	1	1	0	1
14	Subject19	1	1	1	1	1
15	Subject20	1	0	0	0	0
16	Subject21	1	1	1	1	1
17	Subject22	1	1	1	1	1
18	Subject23	1	1	1	1	0
19	Subject24	1	1	1	1	1
20	Subject25	1	1	0	0	0
21	Subject26	1	1	1	0	1

Perinasal Perspiration Data

Here is the list of missing data for perinasal perspiration recording.

Table 12: Data availability of perinasal perspiration data recording, 1 indicates that data is available, 0 indicates data is not available.

	Subject	Session.1	Session.2	Session.3	Session.4	Session.5
1	Subject1	1	1	1	1	1
2	Subject2	1	1	1	1	1
3	Subject3	1	1	1	1	1
4	Subject4	1	1	1	1	1
5	Subject5	1	1	1	0	0
6	Subject6	1	0	0	0	0
7	Subject7	1	1	1	1	1
8	Subject8	1	1	1	1	1
9	Subject9	1	1	0	0	0
10	Subject10	1	1	1	1	1
11	Subject11	1	1	1	1	1
12	Subject12	1	1	1	1	1
13	Subject13	1	1	1	1	1
14	Subject19	1	1	1	1	1
15	Subject20	1	1	1	1	1
16	Subject21	1	1	1	1	1
17	Subject22	1	1	1	1	1
18	Subject23	1	1	1	0	0
19	Subject24	1	1	1	1	1
20	Subject25	1	1	0	0	0
21	Subject26	1	1	1	1	1

6.3 Age Change Information

Below are the users whose age changed during the study period:

- Subject1
- Subject4
- Subject20
- Subject22
- Subject24

7 References

Fast by Nature - How Stress Patterns Define Human Experience and Performance in Dexterous Tasks
https://www.nature.com/articles/srep00305
http://www.sthda.com/english/wiki/ggplot2-line-plot-quick-start-guide-r-software-and-data-visualization
http://r-statistics.co/Linear-Regression.html