STATISTICAL INFERENCE AND QUALITY CONTROL FOR MICROSURGERY STUDY

Swetcha Chowdary Karnati

1 INTRODUCTION

Microsurgery is a general term for surgery requiring an operating microscope. Microsurgical techniques are utilized by several specialties today, such as: general surgery, ophthalmology, orthopedic surgery etc. Microsurgeries require high concentration and preciseness. Any level of stress, deviation or distraction will have serious impact on the performance efficiency of the micro surgeons. Therefore, it is important to analyze and visualize the effects of stress of the surgeon on the accuracy of micro surgery performed.

A longitudinal IRB approved study is performed on 22 medical students to study the relationship of sympathetic arousal and skill in learning micro-surgical tasks.The subjects 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 biographic questionnaire, and a trait anxiety inventory. At the end of their last visit they completed a post-study questionnaire.

During the main part of each session, the subjects underwent the following treatments: Baseline: The subjects were relaxing for 5 min, listening to spa music. They were facially recorded by a thermal and visual camera. Cutting: The subjects had to precision cutting in the inanimate simulator. They were facially recorded by a thermal and visual camera. Suturing: The subjects had to perform suturing in the inanimate simulator. They were facially recorded by a thermal and visual camera. After the cutting treatment the subjects had to fill out a NASA-TLX questionnaire. The subjects also filled out a NASA-TLX questionnaire after the Suturing session. The NASA-TLX instrument features five subscales measuring different aspects of the subjects perceptions regarding task difficulty.

Different data such as Biographic, Trait psychometric, State psychometric,Perinasal perspiration and Performance are collected from the subjects in all the sessions and are used for the analysis.Different quality control plots analyzing the data mentioned above are plotted. Linear Models analyzing and summarizing the effect of stress, task and session on the performance scores are created and analyzed.

1.Biographic data:

The traits of the subjects such as the age and gender are important demographic information and play a significant role in statistical inferences of a research. Plots depicting the age and gender distribution among the subjects involved in the research are summarized in the further sections of the report.

2.Trait psychometric data:

The Trait Anxiety Inventory is a commonly used measure of trait and state anxiety. The TAI questionnare has 20 items for assessing trait anxiety and 20 for state anxiety. State anxiety items include: I am tense; I am worried and I feel calm; I feel secure. Trait anxiety items include: I worry too much over something that really doesnt matter and I am content; I am a steady person. The State-Trait Anxiety Inventory is one of the first tests to assess both state and trait anxiety separately. Each type of anxiety has its own scale of 20 different questions that are scored. Scores range from 20 to 80, with higher scores correlating with greater anxiety.

3.State psychometric data:

The subjects were asked to complete the NASA Task Load Index (TLX) after each session of the task. The NASA-TLX instrument features six subscales measuring different aspects of the subjects perceptions regarding task difficulty. The six sub-scales are Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. The NASA-TLX subscales are scored in the range 1-20.

4.Perinasal Perspiration (Stress) Signal Data:

The perinasal perspiration signal data are the stress measurements obtained during the length of the task.The perinasal perspiration signals are extracted through facial imaging. Here, the perinasal perspiration signal data can be taken as one of the explanatory variable in the linear model used in the study.

5.Performance Data:

The performance of a subject is measured in terms of time taken for the task and accuracy scores obtained for each task. The accuarcy score is described with two variables which are the overall accuracy score per task and the number of sutures performed for the suturing task. The scoring of the accuracy of the task performed is done by two different scorers and is differentiated as scorer1 and scorer2 in the plot depiction in the Appendix section.

2 SUMMARIZING PLOTS THAT REVEAL PATTERNS

The quality control plots drawn for individual subjects are used to analyze the behavior of each subject in different sessions and tasks of the research. In order to produce a summarized view of the quality control plots for all the subjects,box plots are drawn for different quality control variables.

2.1 Biographic Data

The summarizing plots for the biographic data are the same as the quality control plots as the histograms drawn for age and gender distribution are self summarized.

The histogram representing the Gender distribution of the subjects involved in the research is shown below:

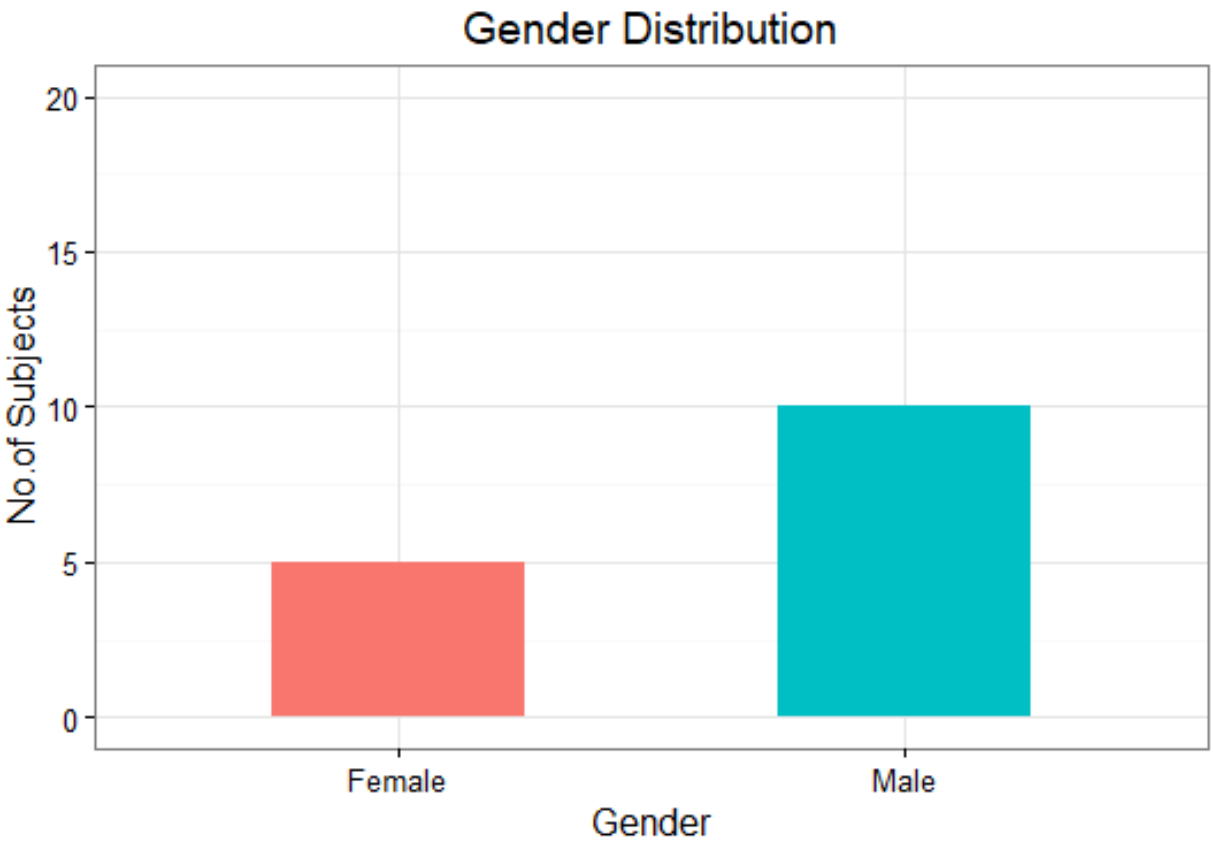


Figure 1: Plot depicting the Gender Distribution of the subjects

It can be observed from the above plot that there are 5 female and 10 male subjects involved in the research. Since, the traits of the subjects play a significant role in statistical inferences of a research, the variable of gender is included in the statistical model constructed for the analysis.

The plots depicting the Age distribution for the subjects involved in the research is shown below:

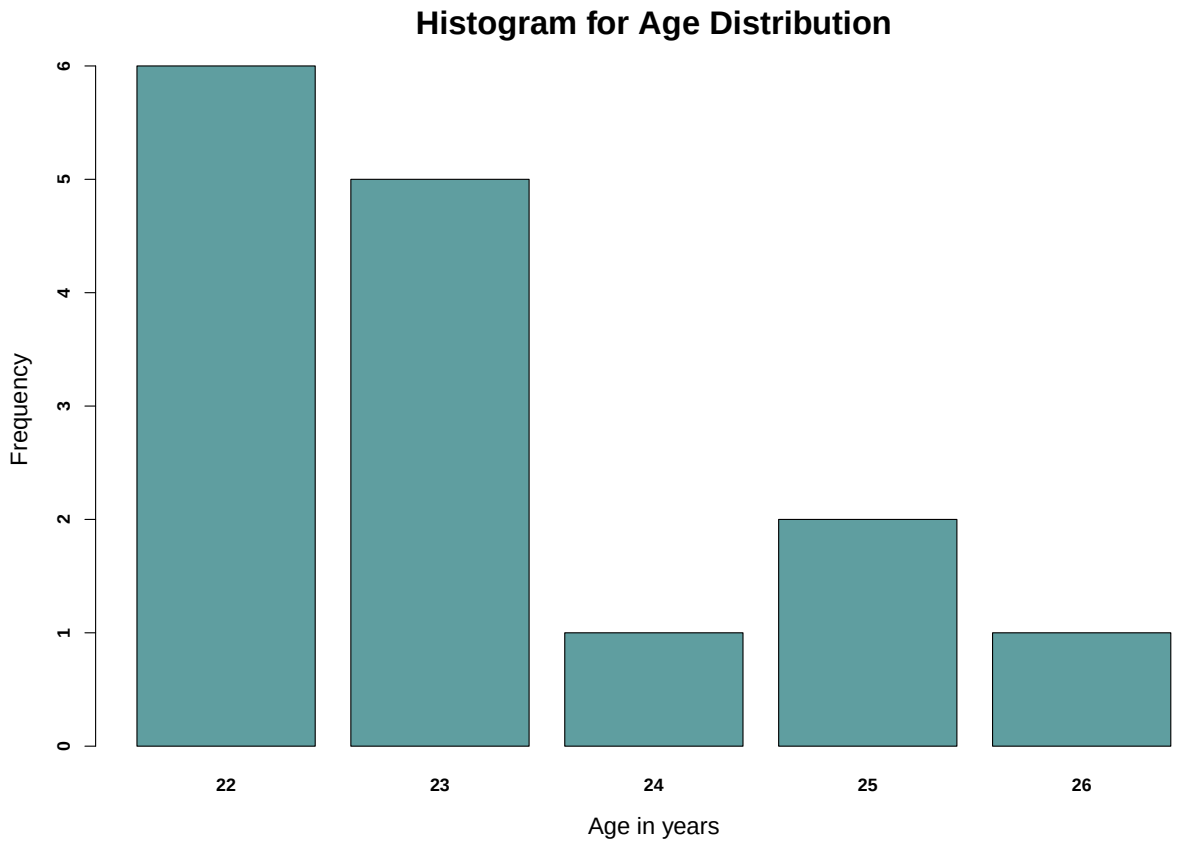


Figure 2: Plot depicting the Age Distribution of the subjects

It can be observed from the plot above that there are six subjects of age 22, five subjects of age 23, one subjects each of age 24 and 26, two subjects of age 25. The average age of the subjects involved in the research is 23.

The variable of age is also included as an independent variable in the model created for further analysis.

2.2 Trait Psychometric Data

The plot depicting the histogram for TAI scores is shown below:

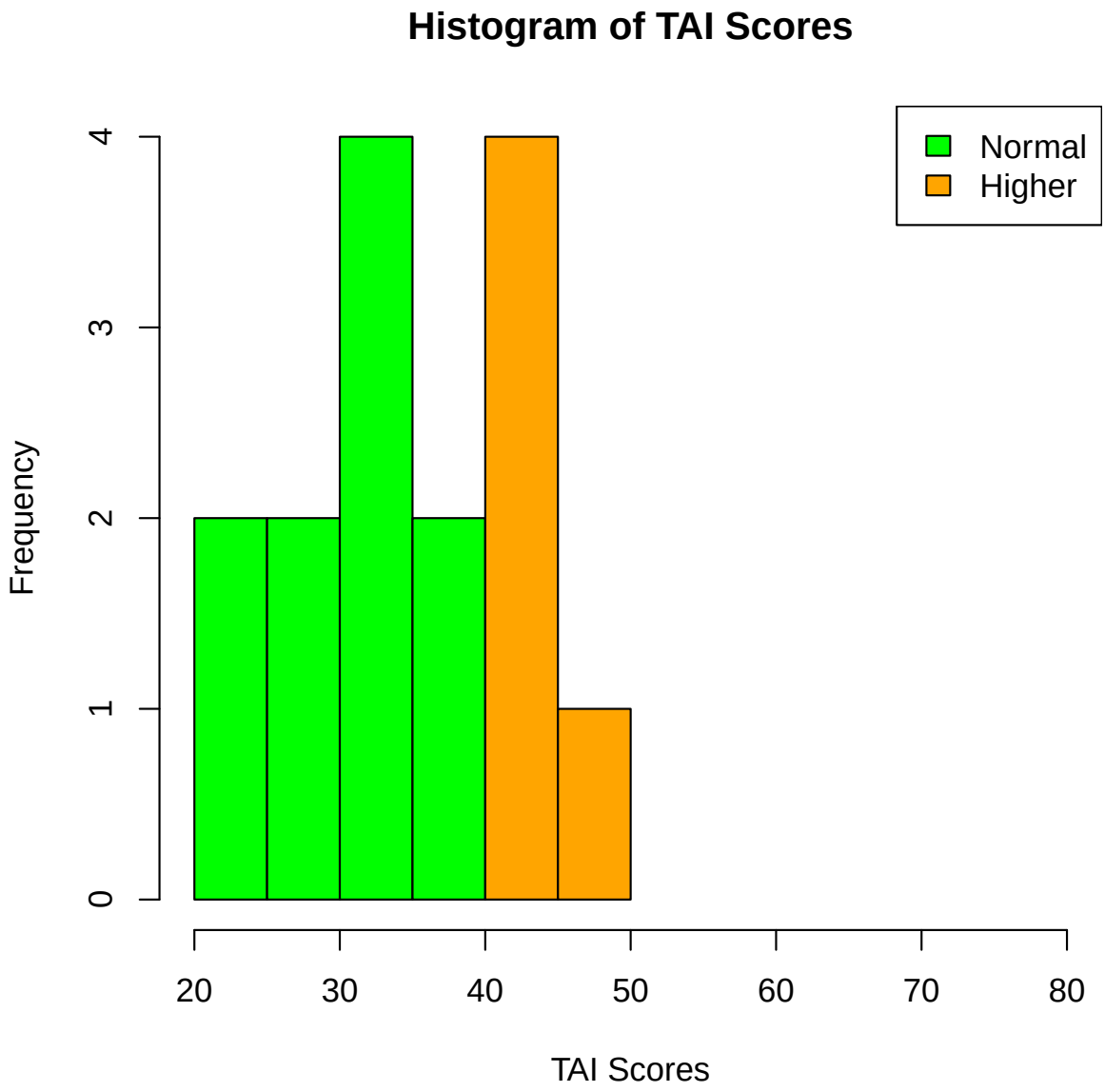


Figure 3: Plot depicting the TAI score Distribution for the 15 subjects who participated in all the sessions

It can be observed from the above histogram that there are 10 subjects out of 15 indicated by green bars in the histogram with TAI scores less than 40 and is considered normal.There are 5 subjects indicated by orange bars in the histogram with TAI scores greater than 40. The higher TAI scores are an indication of greater anxiety.Since the level of anxiety plays an important role in the subject behavior and performance, it is important to analyze the distribution of TAI scores for the subjects.

The boxplot depicting the distribution of TAI scores for different subjects is shown below:

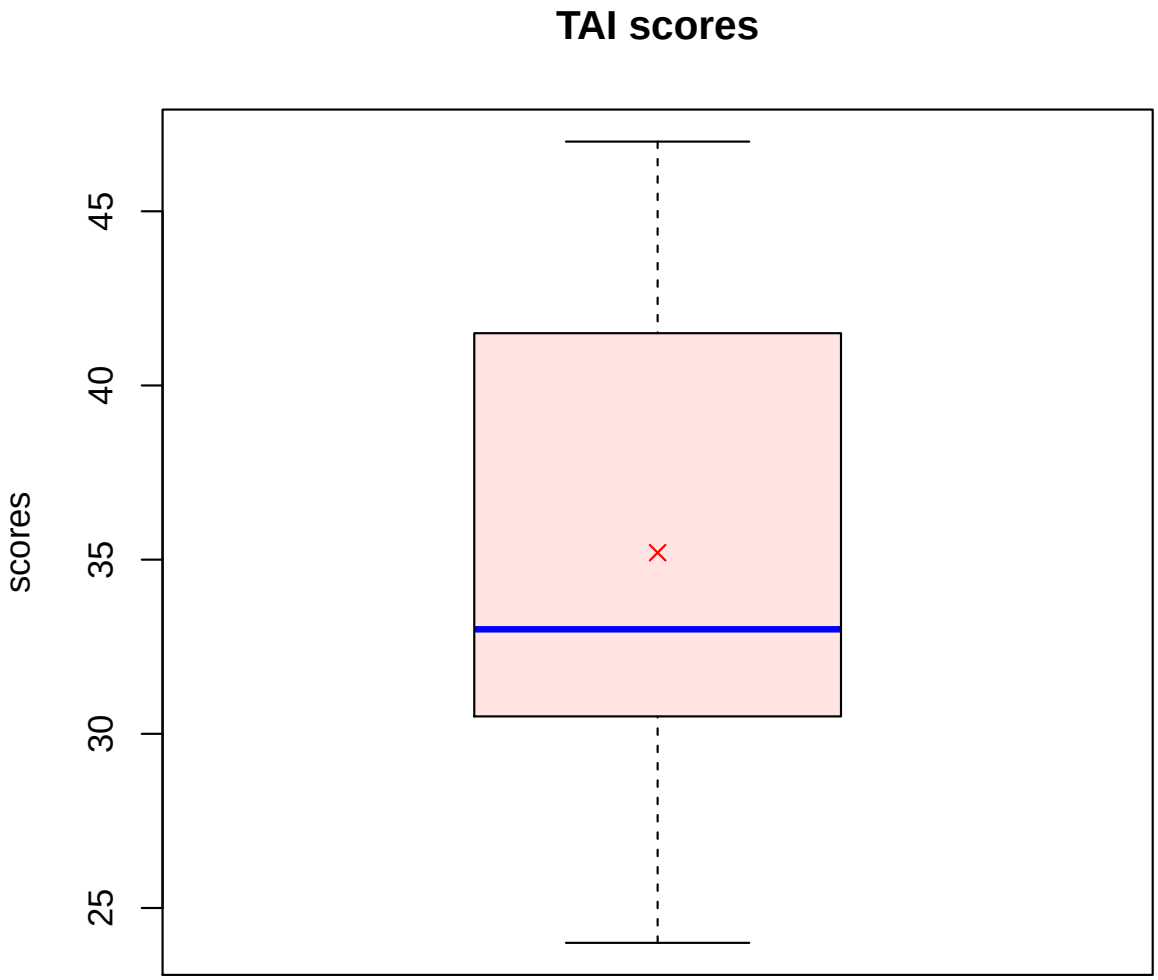


Figure 4: Plot depicting the distribution of TAI scores

It can be observed from the above plots, that the mean of the TAI scores is obtained as 35.2 and standard deviation is 7.0932. It can be observed that the mean of the TAI scores of 35.2 is less than 40 which is considered normal w.r.t the scale of TAI scores.

2.3 State Psychometric data

Plot depicting the NASA TLX score(all sub scales) distribution of cutting task for the sessions is shown below:

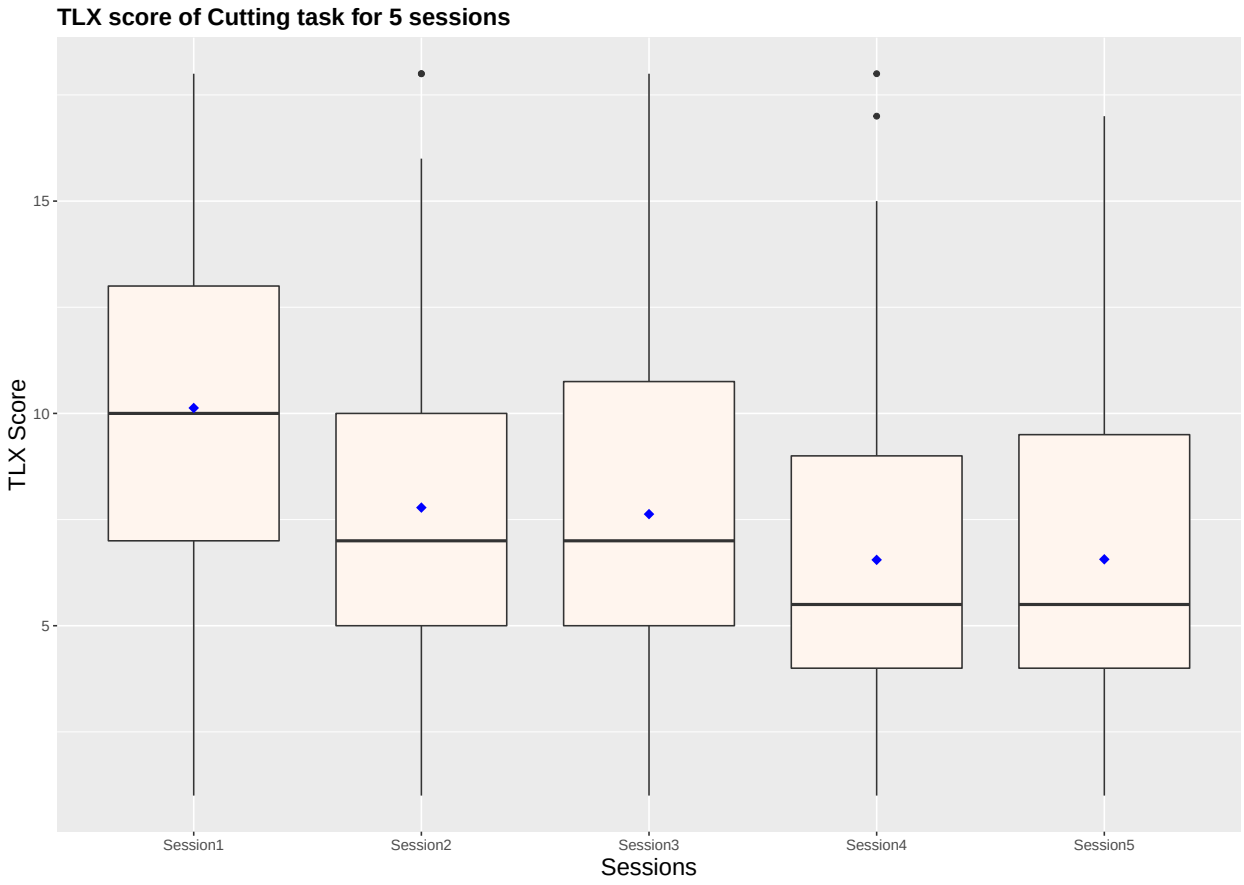


Figure 5: Plot depicting the NASA TLX score distribution of cutting task for the sessions

It can be observed from the above plot that there is a slight decrease in the TLX scores from session 1 to 3, where as scores remained fairly constant in session 4 and session 5.

Further detailed visualization of the distribution of the NASA TLX scores of the subjects for different subscales for the cutting task is shown below:

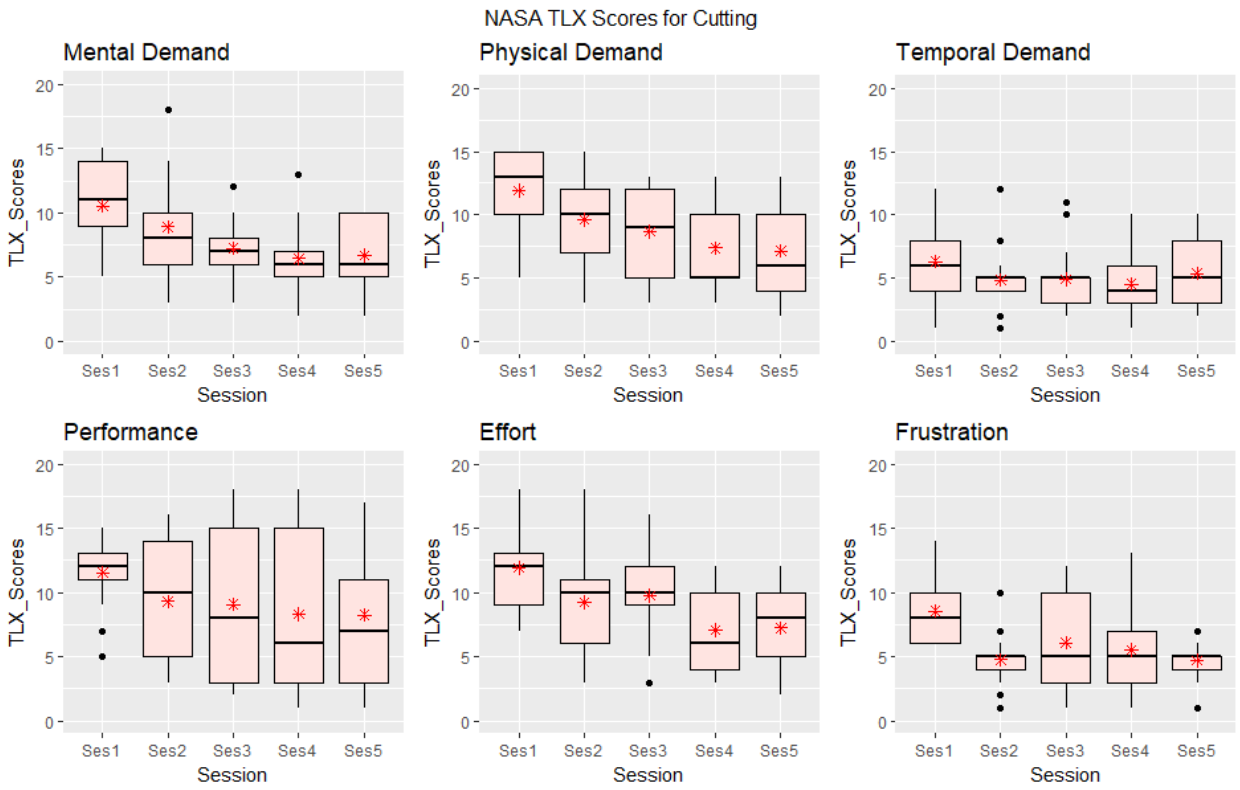


Figure 6: Plot depicting the NASA TLX scores distribution of cutting task for different subscales

Mental Demand for cutting: It can be observed from the boxplot of the mental demand that the mean value of the TLX score distribution gradually decreased from session 1 to session 3 and remained fairly constant after. The shorter box plots in the session3 and session 4 indicate that the subjects have similar TLX scores i.e subjects are in high level of agreement for these sessions. Interestingly the box plot for session 5 indicates higher variability in the quartile group 3 and more agreement of scores in quartile group1.

Physical Demand for cutting: It can be observed from the boxplot of the physical demand that the mean value of the TLX score distribution gradually decreased from session 1 to session 3 and remained fairly constant after. The physical demand box plots show constant reduction in the TLX scores for physical demand.

Temporal Demand for cutting: It can be observed from the box plots that there is a decrease in the mean value of the Temporal demand scores from session 1 to session 3. There are outliers present in the session 2 of the boxplots.The boxplots for temporal demand suggest that there is a slight decrease in the temporal demand from session1 to session 3 but no significant change after session 4.

Performance: Interestingly in the first session of the cutting task, the subjects TLX NASA performance score is high and less varied compared to other sessions. Session 3 has more varied performance TLX score. The mean value of the scores remained fairly constant from

session 2 to session 4.

Effort: It can be observed from the box plots for effort that there is a decrease in the level of NASA TLX effort score from session 1 to session 5. This indicates that the subjects feel more comfortable with the cutting tasks and indicate less effort is required for the tasks as the session increases.The upper whiskers are high for the initial sessions indicating that effort scores are more varied in the region.

Frustration: The NASA TLX scores of frustration indicate interesting patterns. There is no constant decrease or increase in the scores of Frustration. The frustration scores for the session 5 are low and the box is small indicating that the scores are in agreement with less variation among subjects.

Plot depicting the NASA TLX score(all sub scales) distribution of suturing task for the sessions is shown below:

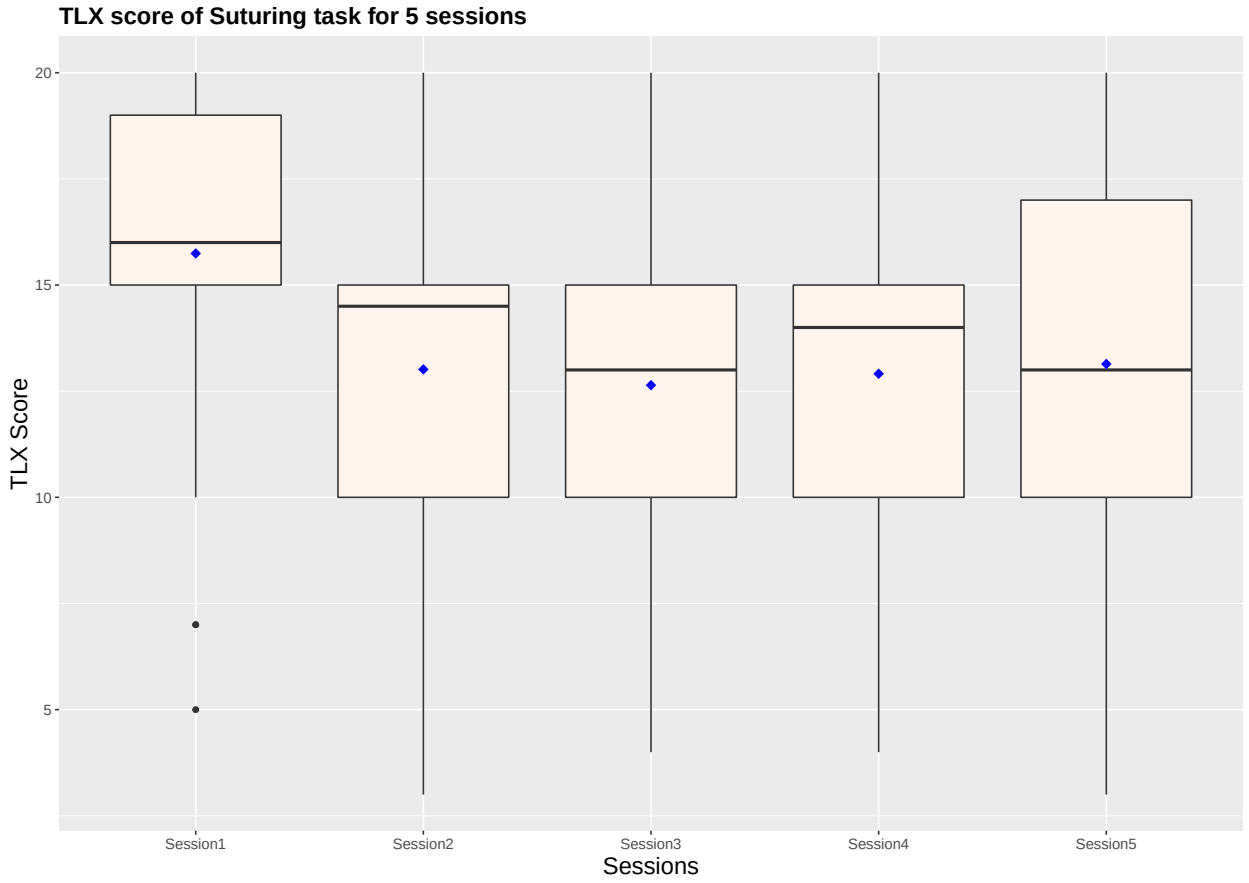


Figure 7: Plot depicting the NASA TLX score distribution of suturing task for the sessions

It can be observed from the above plot that there is a very slight decrease in the TLX scores from session 1 to 3. Interestingly there is a slight increase in the NASA TLX scores in the fourth and fifth sessions. This observation is different compared to the cutting task. In the cutting task, the TLX scores remained fairly constant after session 3 where as in suturing

these scores increased slightly after session 3.

Further detailed visualization of the distribution of the NASA TLX scores of the subjects for different subscales for the cutting task is shown below:

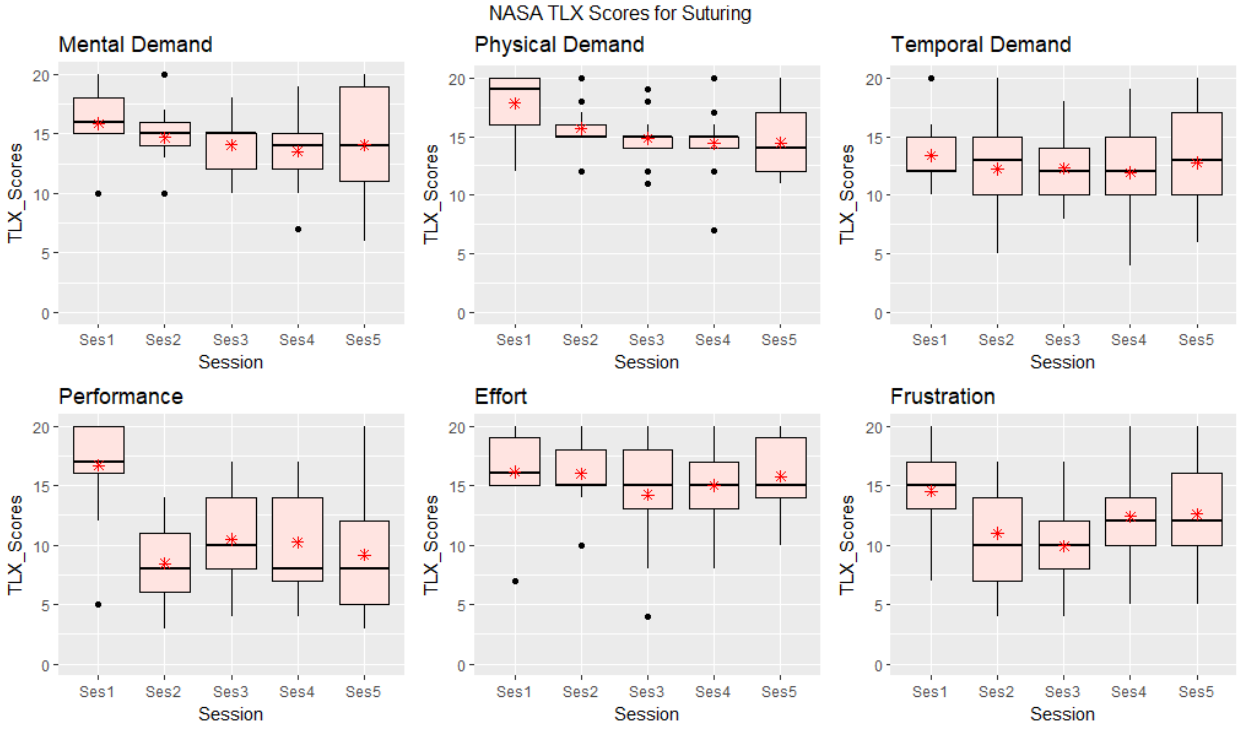


Figure 8: Plot depicting the NASA TLX scores distribution of suturing task for different subscales

Mental Demand for cutting: It can be observed from the boxplot of the mental demand that the mean value of the TLX score distribution gradually decreased from session 1 to session 3 and remained fairly constant after. The shorter box plots in the session2 indicates that the subjects have similar TLX scores of mental demand i.e subjects are in high level of agreement for this session. Interestingly the box plot for session 5 indicates higher variability with higher mental demand scores. The scores are expected to decrease with the increase in sessions but no such observation can be made in the plot.

Physical Demand for cutting: It can be observed from the boxplot of the physical demand that the mean value of the TLX score distribution gradually decreased from session 1 to session 3 and remained fairly constant after. The shorter box plots in the session2, session3 and session 4 indicate that the subjects have similar TLX scores i.e subjects are in high level of agreement for these sessions. Interestingly the box plot for session 5 indicates higher variability scores with more values in upper quartiles.

Temporal Demand for cutting: It can be observed from the box plots that there is a decrease in the mean value of the Temporal demand scores from session 1 to session 2 after which it remained fairly constant.

Performance: Interestingly in the first session of the suturing task, the subjects TLX NASA performance score is high and less varied compared to other sessions. There is a sudden drop in the performance scores from session 1 to session 2. A similar pattern is observed in the cutting task as well. This indicates that the subjects initially believed that they performed well in the first session of task, after which they might have received a feedback.

Effort: It can be observed from the box plots for effort that there no major change in the level of effort required for the task of suturing in different sessions.

Frustration: The NASA TLX scores of frustration indicate interesting patterns. There is a decrease in the frustration from session 1 to session 3 after which it increased with high variability in session5. This indicates that the subjects experienced high frustration in the session1 , session4 and session5 compared to session 2 and session 3.

The boxplot depicting the comparison of the TLX scores distribution for the cutting and suturing tasks is shown below:

Figure 9: Plot depicting the NASA TLX score distribution for Cutting and Suturing tasks

It can be observed from the above plot that the NASA TLX scores (for all subscales and sessions) is less for the cutting task compared to the suturing task. This indicates that the subjects feel that the Mental Demand, Physical Demand, Effort, Frustration, Performance and Temporal Demand required for the cutting task is low when compared to the suturing task. Interestingly 75% of the TLX scores for cutting task are less than the lower quartile of the suturing task. This indicates a possible higher level of difficulty involved in the suturing task compared to the cutting task.

2.4 Perinasal Perspiration (Stress) Signal Data:

The mean perinasal signal values for all the subjects for all sessions are plotted separately the tasks of cutting, suturing and baseline and are shown below :

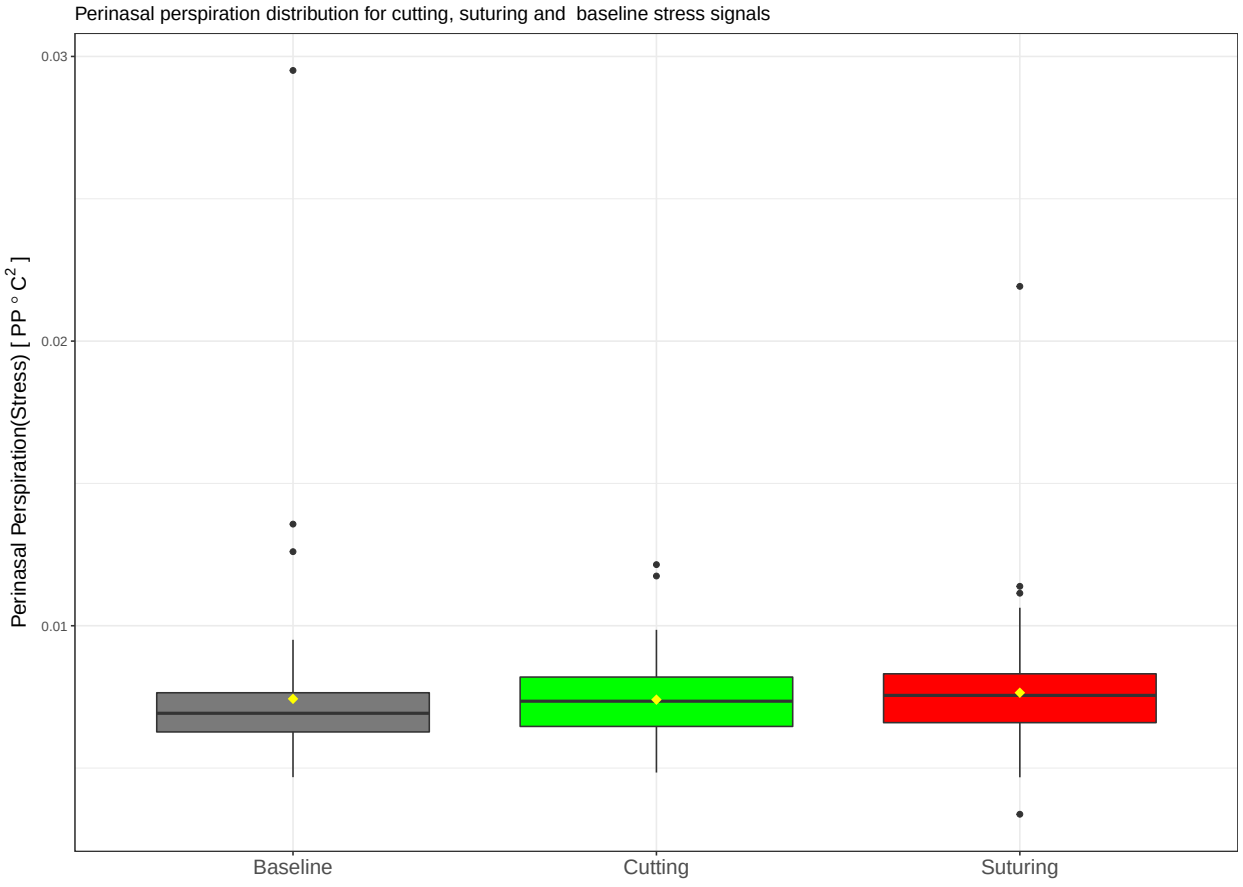


Figure 10: Plot depicting the perinasal perspiration for cutting, suturing and baseline

It can be observed from the above plots that there is no major difference in the values of the perinasal perspiration for cutting, suturing and baseline tasks. It is expected that the subjects would be relieved during the baseline recordings since the subjects would be listening to music during this time. But no such difference in the level of stress is seen in the summarizing boxplots. Further analysis of the dependence of the performance scores on the stress signals is done in the linear model created to analyze the dependency.

2.5 Performance data

The box plot depicting the distribution of performance scores for male and female participants is shown below:

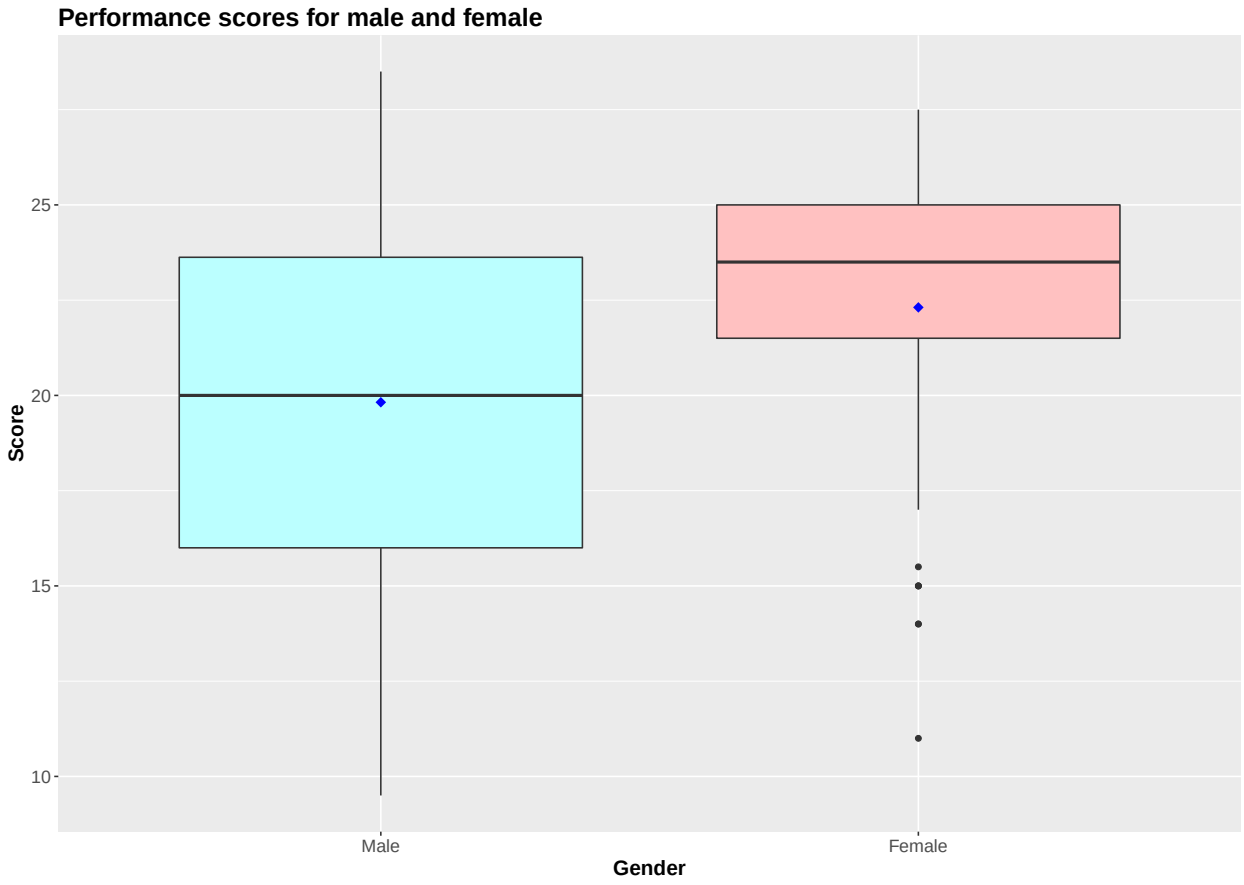


Figure 11: Plot depicting the performance score distribution for and Male and Female participants

In the above plot, it can be observed that the box plot for male is larger in size indicating a varied distribution of performance scores among the male subjects. The box plot for the female scores is comparatively smaller in size indicating that the scores of the female subjects are less varied and more in agreement. The mean of the performance scores for the male subjects is less than the mean of the scores of the female subjects.

The plot depicting the distribution of the time taken for the cutting and suturing tasks for different sessions is shown below:

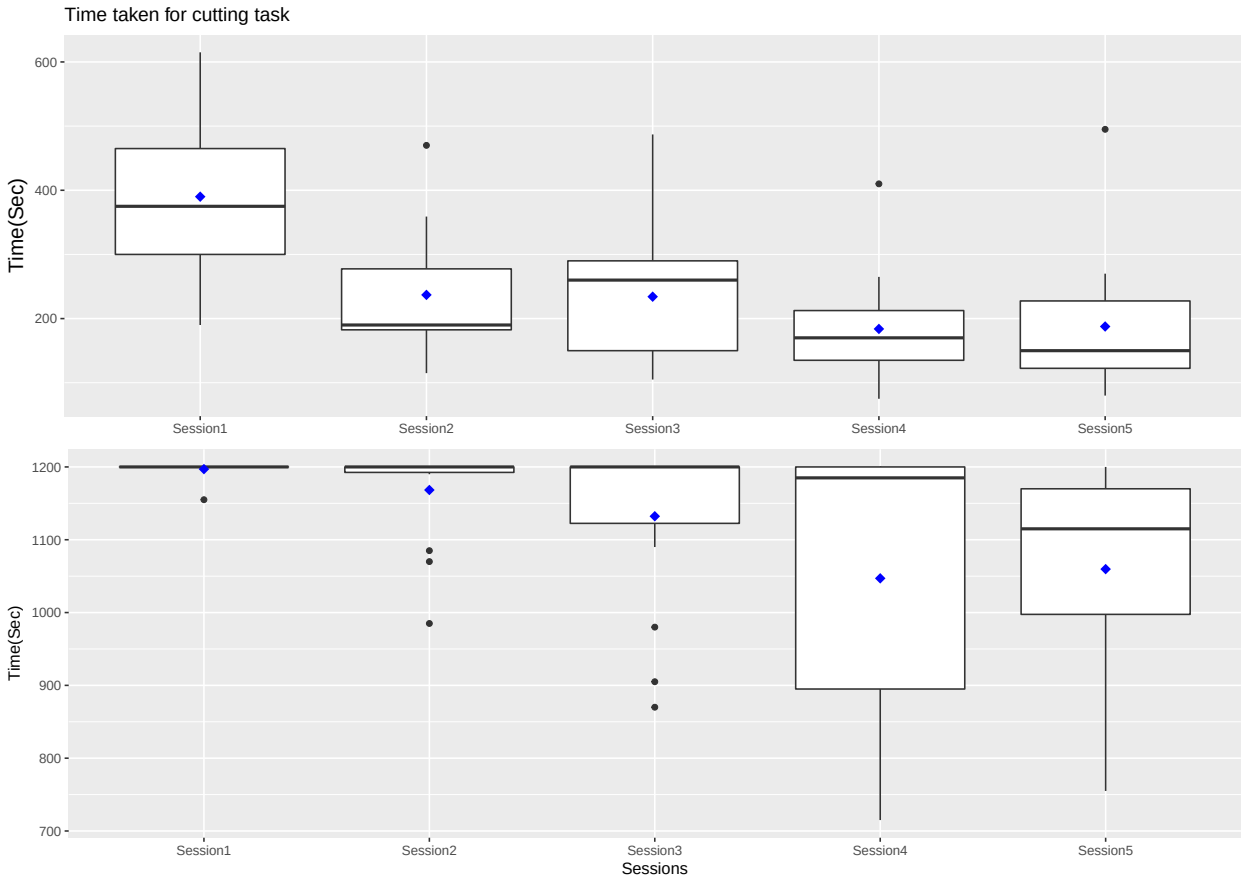


Figure 12: Plot depicting the distribution of the time taken for cutting and suturing tasks

It can be observed from the above plot that there is a significant decrease in the mean of the time taken for the cutting task from session 1 to session 3. The boxplots for the suturing task shows interesting observations with session 1 having a straight line at 1200 seconds indicating that all the subjects have taken approximately 1200 seconds (20 min) to complete the suturing task, where as there is significant decrease in the time taken for suturing task in session 4 and session 5. The long lower whisker in the 4 and 5 sessions indicates that the time taken is varied amongst the most negative quartile group and very similar in the positive quartile group.

This significant decrease from session 1 to 4 indicates that the subjects initially took all the time given (20 min) to finish the suturing task but later on finished the suturing earlier than 20 min time slot.

The boxplot depicting the comparison of the performance scores for the cutting and suturing tasks is given below:

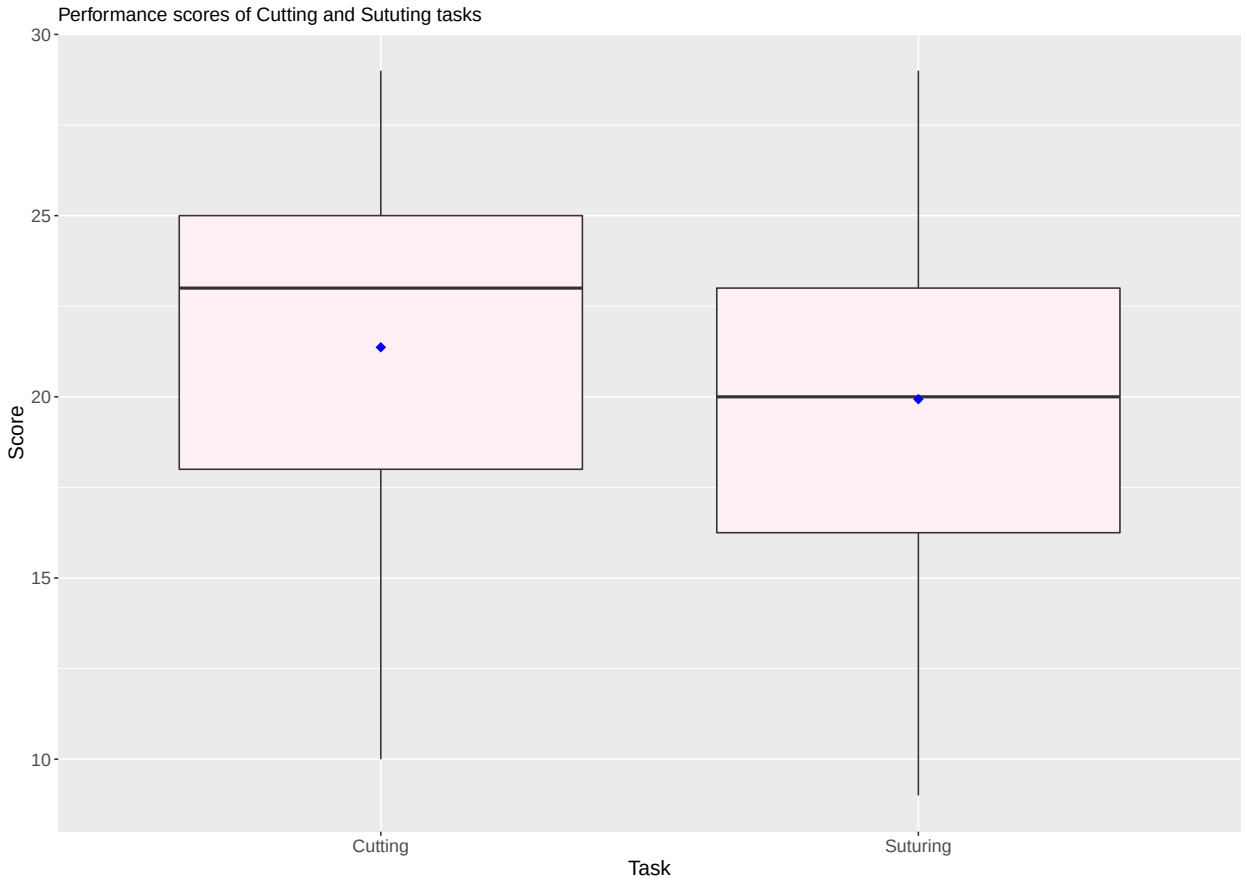


Figure 13: Plot depicting the performance scores distribution for cutting and suturing tasks

It can be observed from the plot that the mean value of the performance scores for the cutting task(21.36) is slightly higher than the mean value of the performance score for the suturing tasks (19.93). The boxplot for the cutting task is uneven with greater variability of scores in the quartile group 2. The boxplots indicate a higher performance score for the cutting task compared to the suturing. This can be verified by including the categorical variable of Type of task in the linear model created to analyze the performance scores.

The box plot depicting the distribution of the performance scores for different sessions for cutting and suturing tasks is shown below:

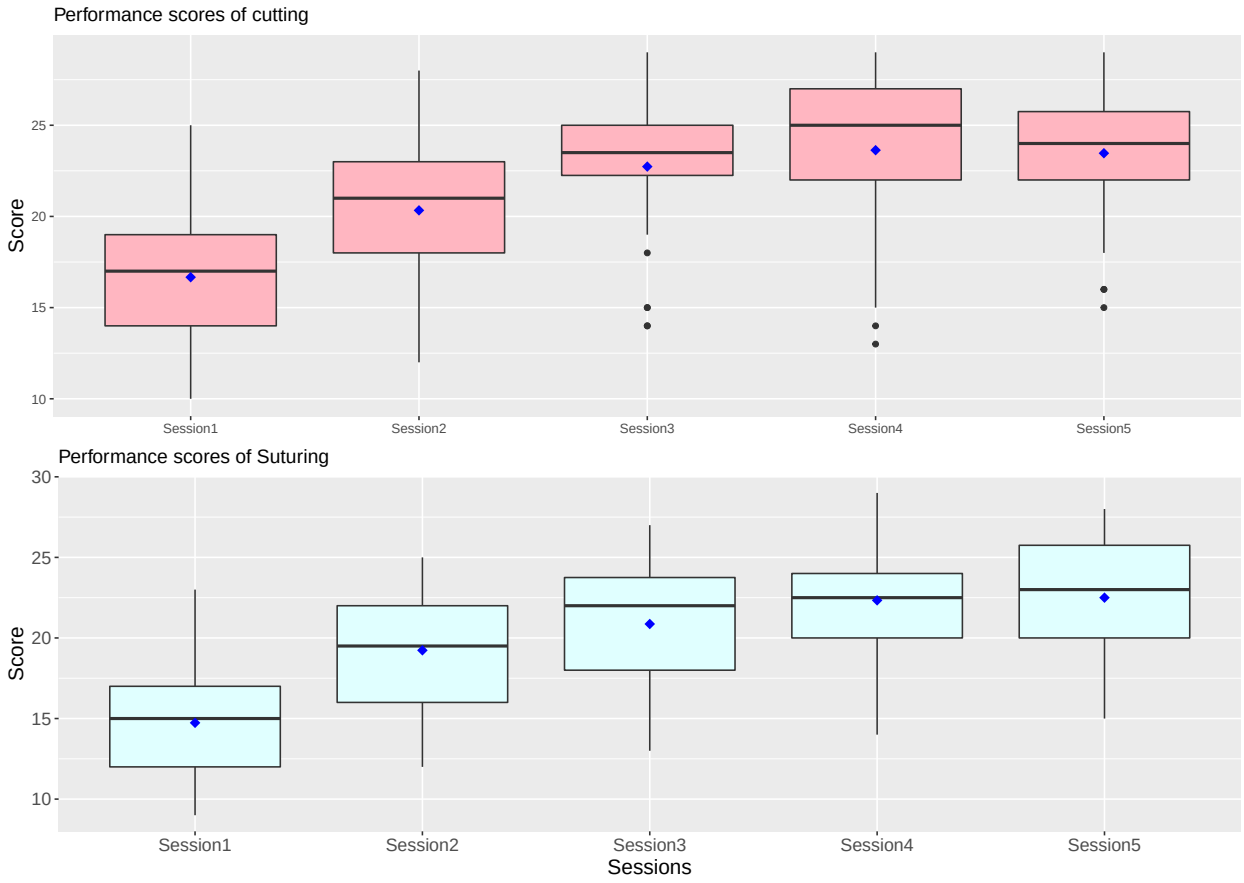


Figure 14: Plot depicting the performance scores distribution for cutting and suturing tasks for all sessions

It can be observed from the above plots that the performance scores increased gradually for the cutting task from session 1 to session 4. The boxplot of cutting performance scores for session 5 is small indicating less variation in the scores of the subjects in this session. This increase in the performance scores is desirable and is expected.

It can be observed from the above plots that the performance scores increased gradually for the suturing task from session 1 to session 5. This increase in the performance scores is desirable and is expected.

3 STATISTICAL INFERENCE

Linear Model:

Linear models describe a continuous response variable as a function of one or more predictor variables. They can help us understand and predict the behavior of complex systems or analyze experimental, financial, and biological data. Here, in order to analyze the effect of stress(perinasal perspiration), session,task, gender,age on the performance score of the

subject, we build two linear models, one without random effects and the other including random effects.

Here, the perinasal perspiration (mupp) is obtained for cutting and suturing tasks by subtracting the mean of the baseline values for each session per subject. The mean perinasal perspiration values obtained are not normalized.

This can be shown by performing a shapiro test as shown below:

```
> shapiro.test(sub_df$mupp)

        Shapiro-Wilk normality test

data:  sub_df$mupp
W = 0.70317, p-value < 2.2e-16
```

The p-value of the shapiro test is less than significance level 0.05 indicating that the mean perinasal perspiration values are not normal. Hence, Linear transformation followed by log transformation is performed on the data.

The linear model is constructed as shown below. The dependent or response variable is the performance scores of the subjects and the independent variables are Sessions, Type of task, Type of scorer, Age and gender of the subject. Sessions, Type of task,type of scorer and gender are taken as categorical variables. The summary of the linear model is shown below:

```
> model1<-lm(scores~log(mupp)+sessions+tasks+scorer+Age+sex,data =
    sub_df)
> summary(model1)

Call:
lm(formula = scores ~ log(mupp) + sessions + tasks + scorer +
    Age + sex, data = sub_df)

Residuals:
    Min      1Q  Median      3Q     Max
-9.5166 -2.2092  0.3482  2.5541  8.4502

Coefficients:
                 Estimate Std. Error t value Pr(>|t|)
(Intercept)       -0.4192    22.2717  -0.019  0.98500
log(mupp)         -4.0289     9.9548  -0.405  0.68599
sessionsSession2   4.1273     0.6707   6.154 2.61e-09 ***
sessionsSession3   6.1565     0.6731   9.146  < 2e-16 ***
sessionsSession4   7.4629     0.6812  10.955  < 2e-16 ***
sessionsSession5   7.4827     0.6665  11.226  < 2e-16 ***
tasksSuturing     -1.3802     0.4261  -3.239  0.00134 **
scorerScorer2      0.1241     0.4236   0.293  0.76968
```

```
Age                0.4111     0.1745    2.356  0.01916 *
sexMale           -2.1222     0.4628   -4.585 6.85e-06 ***
---

Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
      0.1          1

Residual standard error: 3.606 on 280 degrees of freedom
  (10 observations deleted due to missingness)
Multiple R-squared:  0.4444,    Adjusted R-squared:  0.4265
F-statistic: 24.88 on 9 and 280 DF,  p-value: < 2.2e-16
```

Response Variable = Score , Predictor Variables = mupp, sessions, tasks, scorer, age, sex

Interpretation of the Results:

The residuals vs fitted values plot for the model defined above is shown below:

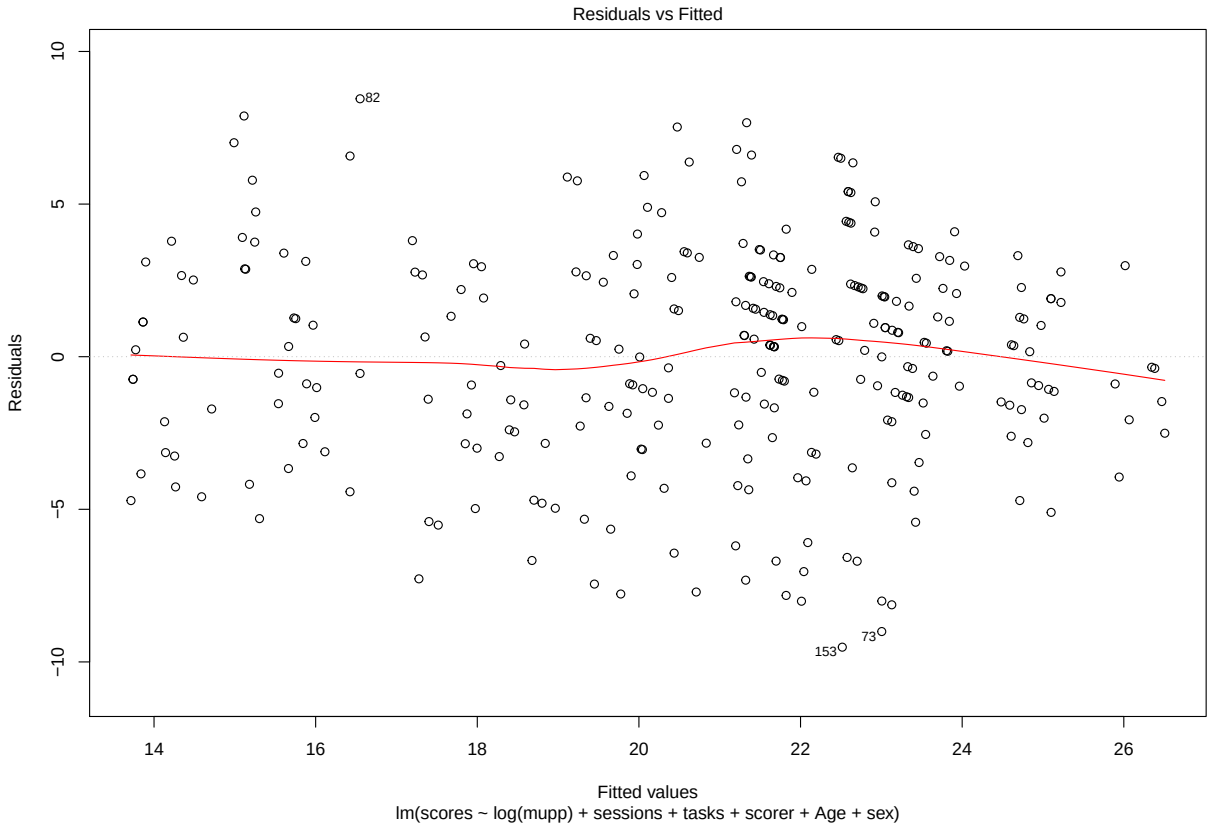


Figure 15: Plot depicting the Residuals Vs Fitted values for the Linear model 1

Residuals: Residuals are essentially the difference between the actual observed response

values and the response values that the model predicted. A residual plot is a graph that shows the residuals on the vertical axis and the independent variable on the horizontal axis. If the points in a residual plot are randomly dispersed around the horizontal axis, a linear regression model is appropriate for the data; otherwise, a non-linear model is more appropriate.

In the above plot for the residuals we can observe that the points are randomly dispersed around the horizontal axis indicating that the linear model used is appropriate for the data. The random dispersion of the points indicates that the model has captured the data correctly and no important information is lost into the residuals.

The Coefficient Estimate shown in the model summary represent the mean change in the response variable for one unit of change in the predictor variable while holding other predictors in the model constant. In the above linear model, we obtained some of the coefficient estimates as negative which means that one unit change in that particular predictor variable by keeping all others constant, will have a negative change or decrease in the response variable by that factor.

The p-values for each independent variable are shown in the model summary. A low p-value less than 0.05 indicates that there is a significant relationship between the dependent and independent variables in the model

The p-value for the model is less than 2.2 e-16 which indicates that there is a significant dependence of the response variable on the independent variables.

R-squared is a statistical measure of how close the data are to the fitted regression line. The R-squared value obtained is 0.444.

Let us consider each predictor variable and their insights:

The log(mupp) value of estimate is -4.0289 which is nothing but Score = (-4.0289)* (mupp) (By keeping the remaining predictor variables as constant). Which indicates that the unit change in mupp value decreases the score by a factor of -4.0289 when the remaining predictor variables are kept constant. The p-value for the variable of log(mupp) is 0.68599 which is greater than significance level of 0.05 indicating that the variable of log(mupp) or stress has no impact on the response variable of performance score. This indicates that the stress values do not impact the performance scores of the subject.

The sex Male estimate is -2.1222 which indicates that the Female subjects have better performance scores when compared to the Male subjects. The significance of three stars and a p-value of 6.85e-06 indicates that the variable sex has significant impact on the response variable of performance scores with an indication of better female performance scores than male. The same is also obtained in the box plot of section 2 which shows the comparison of performance scores based on gender.

The Scorer variable has a p-value of 0.76968 and no significance. This shows that the type of scorer does not play a role in determining the scores obtained which shows that the scorer1 and scorer2 are in agreement and scores by scorer 1 and 2 are almost equal. This indicates that the parameters considered by both the scorers in the scoring process are equal and there is no bias on behalf of the scorers.

The variable task has significance of two stars and a p-value of 0.00134 indicating that the type of task plays an extremely significant role in determination of the performance score.The task suturing has an -1.3802 estimate which shows that the task of cutting has better performance scores when compared to the suturing. The inference obtained here is similar to the inference obtained in the plots for the comparison of performance scores of cutting and suturing tasks both indicating that the scores obtained for cutting task are better than suturing task.

The variable Age has p-value = 0.01916 and significance of one-star which shows that the dependence of performance scores on age of the subject is significant but not extremely significant as other variables such as session, task and sex.

All the Sessions factors have three stars for the significance indicating that the dependent variable i.e scores significantly depends on the session variable.The estimate values of the sessions also shows a clear increase from over sessions from session 1 to session 5 indicating that the performance scores of the subjects increased with the increase in the number of session. This is an inherent indication of the known norm that accuracy increases with practice.

This increase is fairly large initially i.e in the sessions 2 and 3 but as can be observed in the output the difference between values of session 4 (7.4629) and session 5 (7.4827) is low(almost constant). The same inference is obtained earlier from the box plots drawn to compare the increase/decrease of performance scores for cutting and suturing tasks in section 2 of the report.

Mixed model:

A mixed model is a statistical model containing both fixed effects and random effects. These models are useful in a wide variety of disciplines in the physical, biological and social sciences.

Here, the performance of the subject not only depends on the variables determined in the earlier linear model. Performance of a subject is also dependent on the subject's inherent characteristics. These characteristics are different for different subjects and are characterized using random effects. In order to consider the random effects along with other independent variables discussed earlier, we design a mixed model by taking the subject as the random variable.

The design and summary of the mixed model is shown below:

```
> model2<-lmer(scores~log(mupp)+sessions+tasks+scorer+Age+sex+(1|
    subjects),data = sub_df)
> summary(model2)
Linear mixed model fit by REML ['lmerMod']
Formula: scores ~ log(mupp) + sessions + tasks + scorer + Age +
   sex +      (1 | subjects)
   Data: sub_df

REML criterion at convergence: 1439

Scaled residuals:
```

```
    Min      1Q  Median      3Q     Max
-2.8485 -0.5150  0.0758  0.6516  2.2901

Random effects:
 Groups   Name        Variance Std.Dev.
 subjects (Intercept) 6.321    2.514
 Residual             7.761    2.786
Number of obs: 290, groups:  subjects, 15

Fixed effects:
                  Estimate Std. Error t value
(Intercept)        -8.9229    22.6651  -0.394
log(mupp)          -8.0403     8.4056  -0.957
sessionsSession2    4.0884     0.5190   7.878
sessionsSession3    6.2128     0.5221  11.899
sessionsSession4    7.4913     0.5283  14.181
sessionsSession5    7.4967     0.5153  14.547
tasksSuturing      -1.2816     0.3297  -3.887
scorerScorer2       0.1241     0.3272   0.379
Age                 0.4055     0.5517   0.735
sexMale            -2.1716     1.4714  -1.476

Correlation of Fixed Effects:
            (Intr) lg(mp) sssnS2 sssnS3 sssnS4 sssnS5 tsksSt
               scrrS2 Age
log(mupp)    0.817
sessnsSssn2 -0.048 -0.043
sessnsSssn3 -0.196 -0.226  0.487
sessnsSssn4 -0.152 -0.177  0.479  0.509
sessnsSssn5 -0.077 -0.083  0.487  0.499  0.494
tasksSutrng -0.098 -0.113  0.005  0.026  0.023  0.022
scorerScrr2 -0.007  0.000  0.000  0.000  0.000  0.000  0.000
Age         -0.604 -0.037  0.005  0.008  0.001  0.000  0.001
   0.000
sexMale     -0.171  0.024  0.004 -0.006 -0.001  0.000 -0.002
   0.000  0.261

> AIC(model2)
[1] 1462.982
> BIC(model2)
[1] 1507.021
```

By looking at the results, we get an estimate of the variance explained by the random effect. This number is important, because if it's indistinguishable from zero, then the random effect probably doesn't matter and we can go ahead and do a regular linear model instead. So,

here we can see that there are variances of 6.327 which is clearly distinguishable from zero. So, there is a necessity to include the Random effect. The creators of the lme4 package are philosophically opposed to p-values, so to get more detailed analysis we can get F-values from the anova function as shown below:

```
> anova(model2)
Analysis of Variance Table
           Df  Sum Sq Mean Sq F value
log(mupp)   1   27.38   27.38  3.5284
sessions    4 2277.73  569.43 73.3713
tasks       1  117.54  117.54 15.1446
scorer      1    1.12    1.12  0.1440
Age         1   10.44   10.44  1.3453
sex         1   16.91   16.91  2.1784
```

Certainly when we look at the fixed effects estimate, we can see that log(mupp) has an estimate coefficient of -8.9229 which indicates that the unit change in log(mupp) value decreases the score by a factor of -8.9229 when the remaining predictor variables are kept constant. Rest of all the predictor variables are all similar to the Linear model interpretations with not of much significant difference.

Penalized likelihood methods calculate the likelihood of the observed data using a particular model. But because it is a fact that the likelihood always goes up when a model gets more complicated, whether or not the additional complication is justified, a model complexity penalty is used. Several different penalized likelihoods are available such as BIC (Bayesian information criterion). AIC (Akaike information criterion) . The BIC number penalizes the likelihood based on both the total number of parameters in a model and the number of subjects studied. The AIC and BIC values are obtained as 1462.982 and 1507.021 respectively.

Model3:

The NASA Task Load Index (NASA-TLX) is a widely used subjective, multidimensional assessment tool that rates perceived workload in order to assess a task, system, or team's effectiveness or other aspects of performance.It is important to analyze the variation of NASA TLX scores for different sessions, tasks and sub scales.

Hence, a linear model is created with NASA TLX scores as the dependent variable and factor variables of session, task and NASA TLX subscale as the independent variable. The summary of the model is shown below:

```
> mod_tlx<-lm(TLX_Scores~Session+Task+NASA_TLX_Type,data =
    score_cut_sut)
> summary(mod_tlx)

Call:
lm(formula = TLX_Scores ~ Session + Task + NASA_TLX_Type, data =
    score_cut_sut)
```

```
Residuals:
     Min       1Q   Median       3Q      Max
-10.6423  -2.5500   0.0949   2.4077  11.6128

Coefficients:
                                 Estimate Std. Error t value Pr(>|t|)
(Intercept)                      11.68462    0.44912  26.017  < 2e-16
   ***
SessionSession2                  -2.53846    0.42822  -5.928 4.63e-09
   ***
SessionSession3                  -2.80128    0.42822  -6.542 1.11e-10
   ***
SessionSession4                  -3.20513    0.42822  -7.485 1.96e-13
   ***
SessionSession5                  -3.08333    0.42822  -7.200 1.43e-12
   ***
TaskSuturing                      5.75897    0.27083  21.264  < 2e-16
   ***
NASA_TLX_TypeFrustration         -3.23077    0.46909  -6.887 1.18e-11
   ***
NASA_TLX_TypeMental Demand       -1.03077    0.46909  -2.197   0.0283
   *
NASA_TLX_TypePerformance         -2.09231    0.46909  -4.460 9.40e-06
   ***
NASA_TLX_TypePhysical Demand     -0.03846    0.46909  -0.082   0.9347
NASA_TLX_TypeTemporal Demand     -3.37692    0.46909  -7.199 1.45e-12
   ***
---
Signif. codes:  0    ***    0.001    **    0.01    *    0.05    .
      0.1          1

Residual standard error: 3.782 on 769 degrees of freedom
Multiple R-squared:  0.4512,    Adjusted R-squared:  0.4441
F-statistic: 63.22 on 10 and 769 DF,  p-value: < 2.2e-16
```

Interpretation of the Results:

The p-value for the model is less than 0.05 indicating that the model is significant i.e there is a significant dependence of the dependent variable on the independent variables.

All the Sessions factors have three stars for the significance indicating that the dependent variable i.e NASA TLX scores significantly depend on the session variable.The estimate values of the sessions also shows a clear decrease over sessions from session 1 to session 4 indicating that the NASA TLX scores of the subjects decreased with the increase in the number of session. The NASA TLX scores remained fairly constant with a slight decrease

from session 4 to session 5. This inference is also obtained in the summarizing plots in section 2.

The task variable has high significance indicating that the type of task plays a significant role in determining the NASA TLX sub scale values. Here it can be observed that the estimate value for the suturing task is positive (5.75897) indicate that the NASA TLX scores for the suturing task are higher compared to the suturing task.

The significance varied for different NASA TLX sub scales with the NASA TLX Type Frustration, Performance and Temporal Demand having the highest level of significance. The sub scales of Mental Demand is just significant where as the Physical Demand has no significance. The estimate values are different for different sub scales but all are negative indicating that these values are less than the Effort sub scale. The more detailed explanation of the variation of the NASA TLX scores is mentioned in the section 2.

Model 4:

A linear model is constructed in order to observe the dependence of the performance scores on the interaction between the tasks and sessions.

```
> model1<-lm(scores~sessions*tasks,data = sub_df)
> summary(model1)

Call:
lm(formula = scores ~ sessions * tasks, data = sub_df)

Residuals:
     Min       1Q   Median       3Q      Max
-10.633  -2.500   0.500   2.567   8.333

Coefficients:
                                Estimate Std. Error t value Pr(>|t
                                   |)
(Intercept)                     16.66667    0.69914  23.839  < 2e
   -16 ***
sessionsSession2                 3.66667    0.98873   3.708
   0.00025 ***
sessionsSession3                 6.06667    0.98873   6.136 2.77e
   -09 ***
sessionsSession4                 6.96667    0.98873   7.046 1.34e
   -11 ***
sessionsSession5                 6.80000    0.98873   6.878 3.75e
   -11 ***
tasksSuturing                   -1.93333    0.98873  -1.955
   0.05150 .
sessionsSession2:tasksSuturing   0.83333    1.39827   0.596
   0.55166
```

```
sessionsSession3:tasksSuturing   0.06667     1.39827    0.048
   0.96201
sessionsSession4:tasksSuturing   0.63333     1.39827    0.453
   0.65093
sessionsSession5:tasksSuturing   0.96667     1.39827    0.691
   0.48991
---
Signif. codes:  0     ***     0.001     **     0.01     *     0.05     .
       0.1          1

Residual standard error: 3.829 on 290 degrees of freedom
Multiple R-squared:  0.3621,     Adjusted R-squared:  0.3423
F-statistic: 18.29 on 9 and 290 DF,  p-value: < 2.2e-16
```

It can be observed from the above model that the interactions between the sessions and task are not significant and hence do not affect the initial linear model constructed. There is no relationship between the performance scores obtained and the interaction of the session and task variables.

4 DISCUSSION

After analyzing the linear models and observing the summarizing and quality control plots,it can be observed that the performance scores of the subjects depend on the variable of sessions with an increase in the performance scores as the session increases. This indicates that the subjects have an improvement in performance of the microsurgical tasks with increase in practice sessions involved. This increase in the performance scores over sessions is observed to be high initially but remained fairly constant after session 4 which shows that there is a maximum performance score which could be reached with practice sessions after which the scores remain fairly constant.

Also, with the increase in the number of sessions it is observed in the Fig 12 that there is a decrease in the time taken for performing the cutting and suturing tasks. This indicates that besides proficiency, speed of the subject in performing the microsurgery also increases with the increase in number of sessions.

We can also observe from the figures 5 , figure 7 and model 3 that there is a decrease in the NASA TLX subscales scores indicating that the subjects feel less physical demand, mental demand, frustration, effort and temporal is required for the task as the lab practice sessions increases.This indicates that the subjects perceive the workload as less of a burden.

The variable of age does not play any significant role in determining the performance scores indicating that the age of the subject is not an important factor in determining the performance of the subjects.

It is observed from the linear model and summarizing plot 13 that the task of suturing has less performance scores compared to the suturing task. This indicates that the subjects have performed well in the cutting task than suturing task. Interestingly from the figure 9 it

can be observed that the NASA TLX scores obtained for the suturing task are higher than the cutting task indicating that the workload as perceived by the subject is higher for the suturing task as compared to the cutting task.

The summarizing plot for the perinasal mean values (figure 10) and the linear model indicates that there is no significant effect of the perinasal perspiration values measured for the subjects on the performance scores.

The summarizing plot of figure 11 and the linear model indicates that the performance scores for the female subjects is higher than the male subjects.

The research indicates that as the practice with the sessions or experience increases, there is an increase in the performance scores and speed of the task performed with a decrease in the work load as perceived by the subject. Therefore it is important to train the micro surgeon students to observe an increase in the performance, accuracy and speed of the surgeries performed.

5 APPENDIX: QUALITY CONTROL PLOTS

5.1 Biographic Data

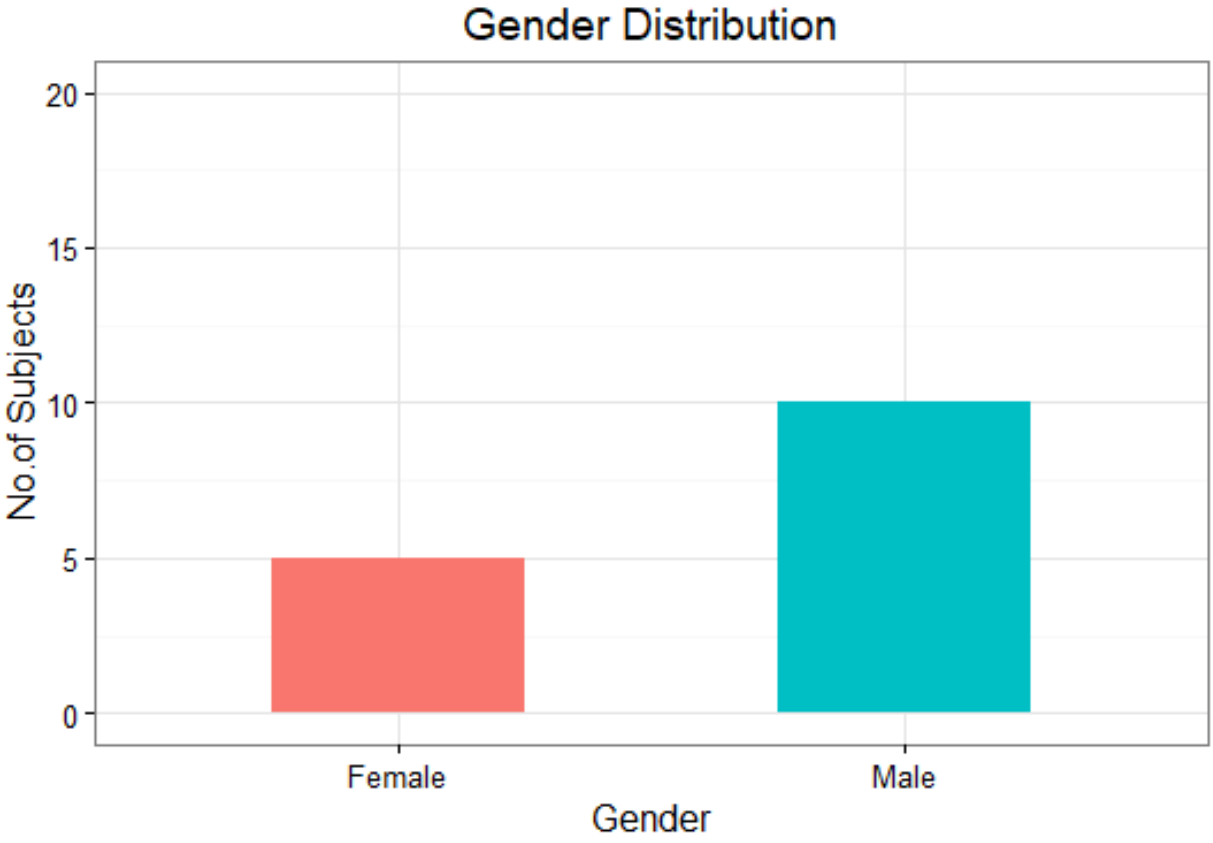


Figure 16: Plot depicting the Gender Distribution of the subjects

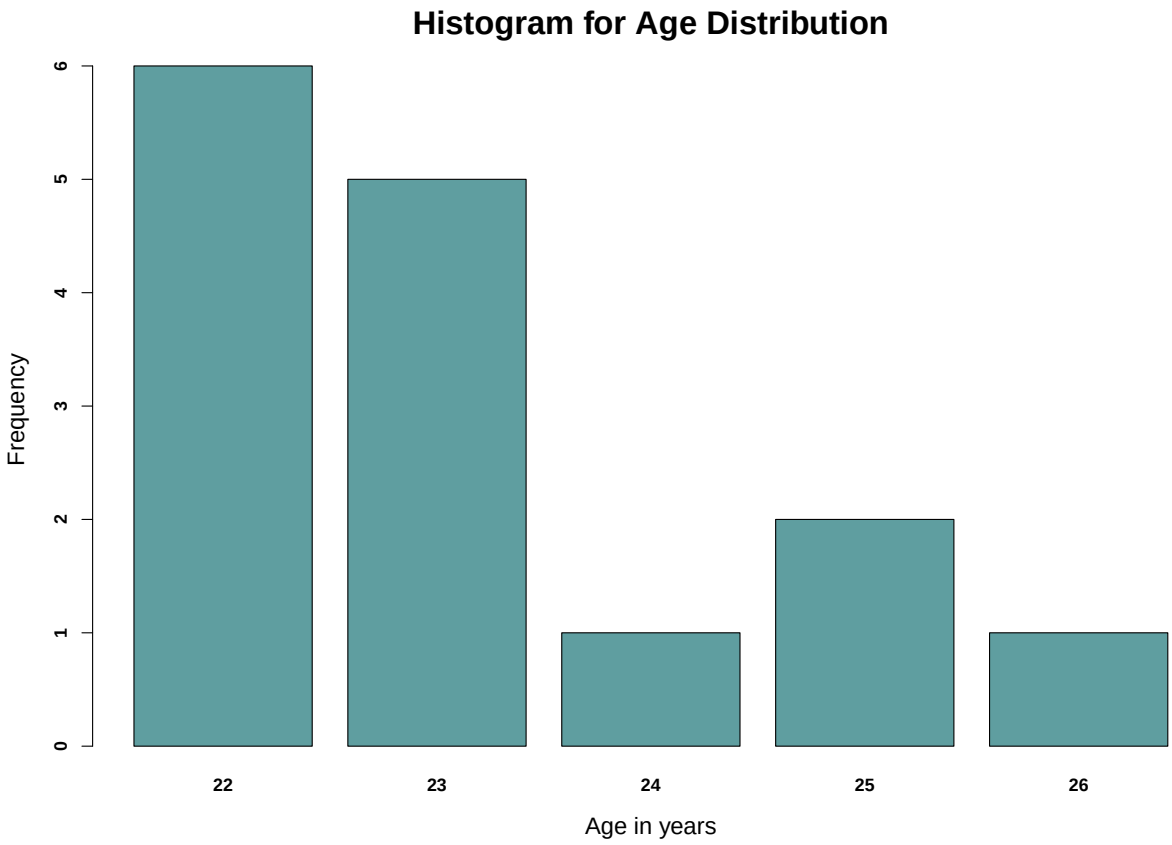


Figure 17: Plot depicting the Age Distribution of the subjects

5.2 Trait Psychometric Data

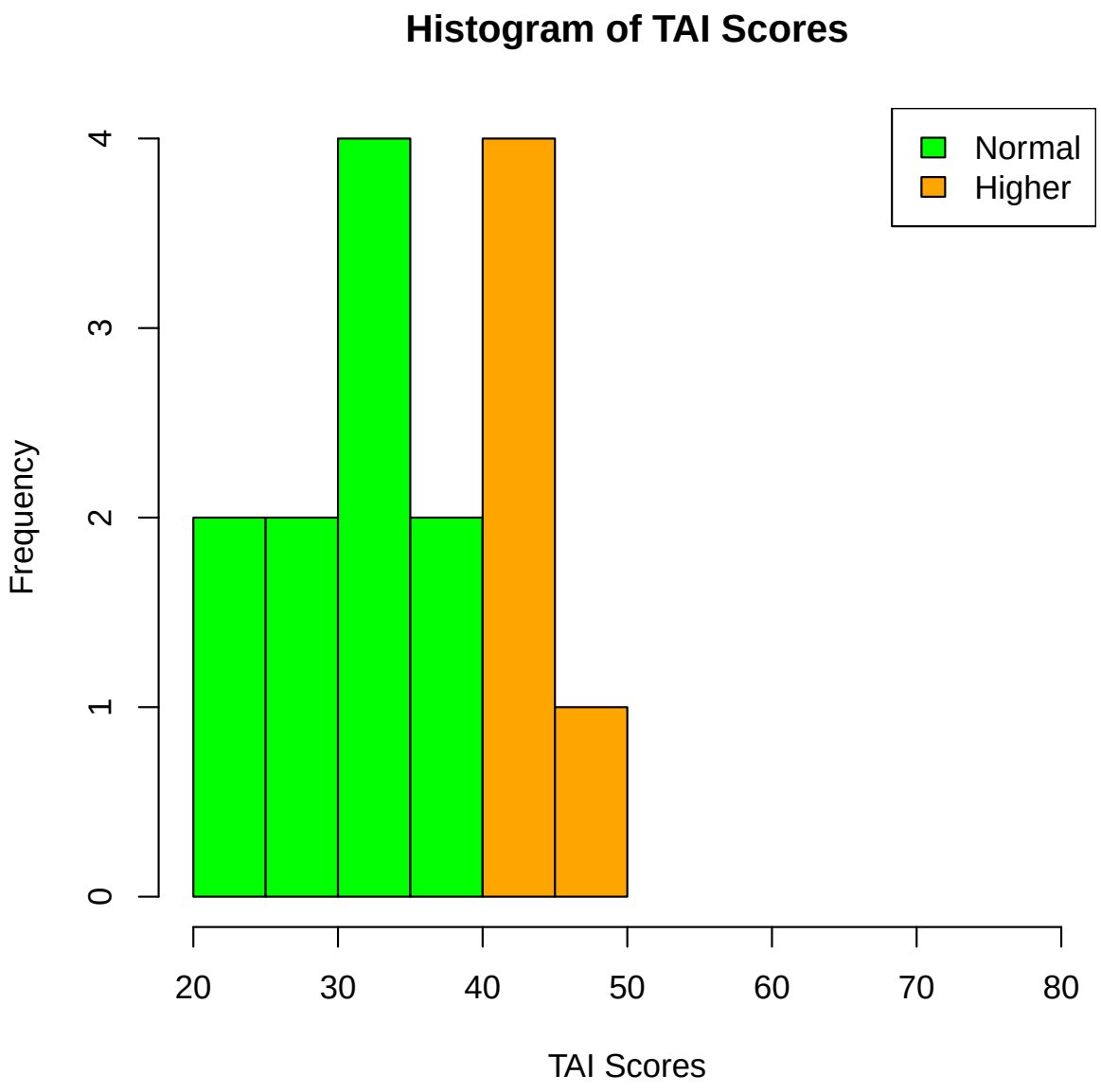
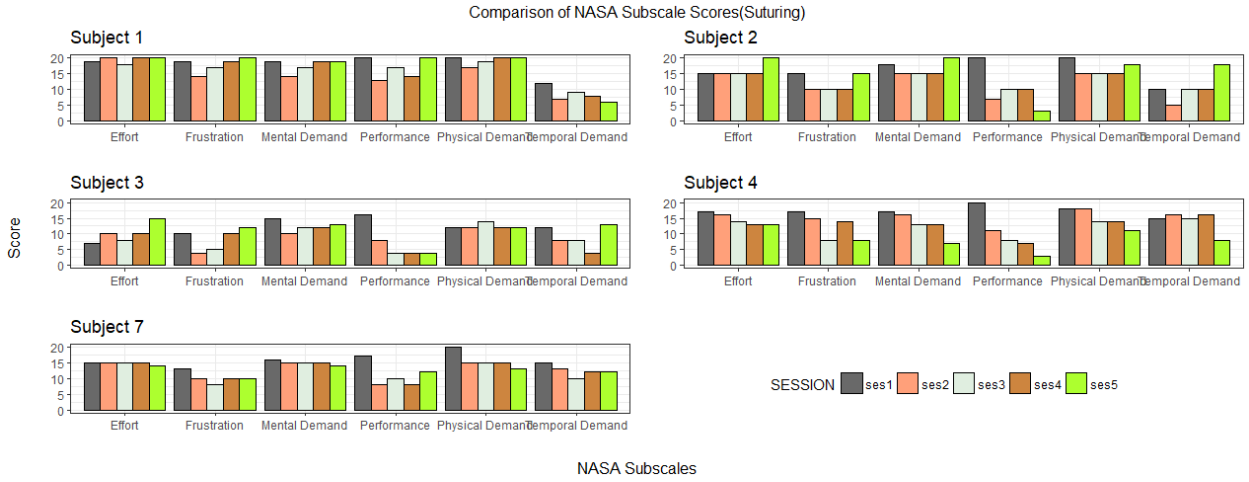
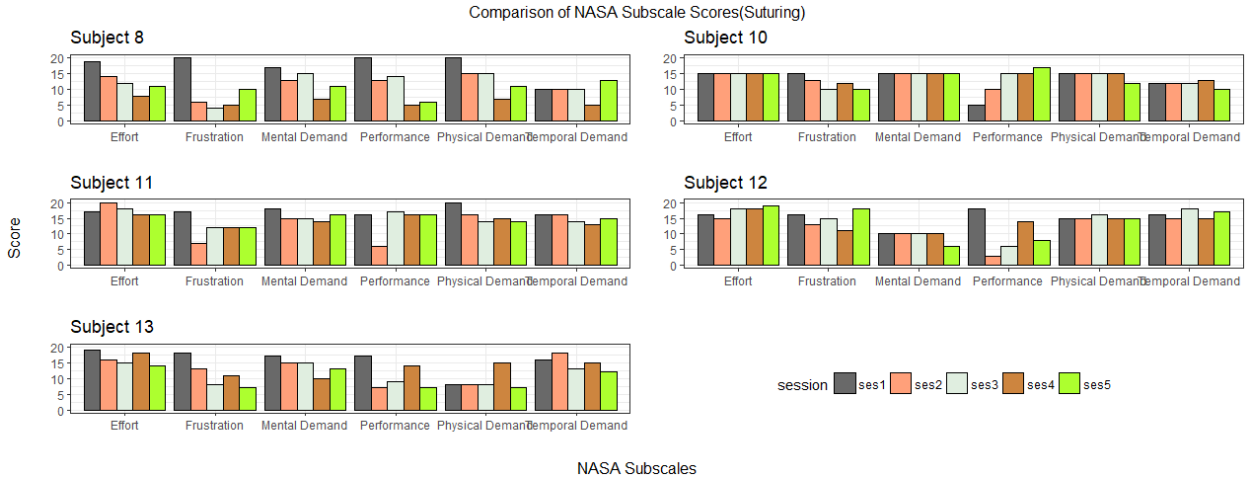


Figure 18: Plot depicting the TAI score Distribution for the 15 subjects who participated in all the sessions

5.3 State Psychometric Data





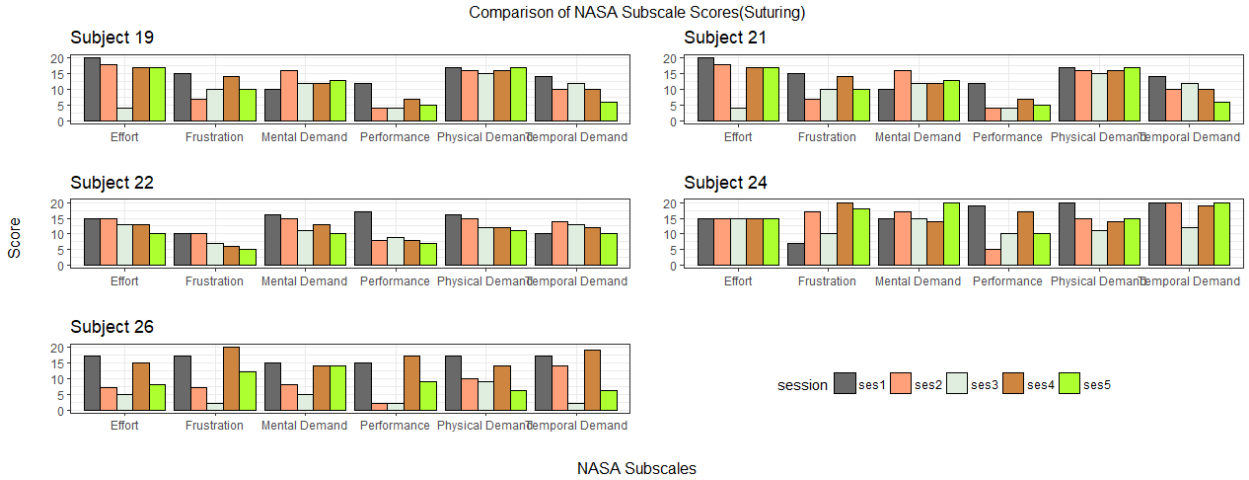
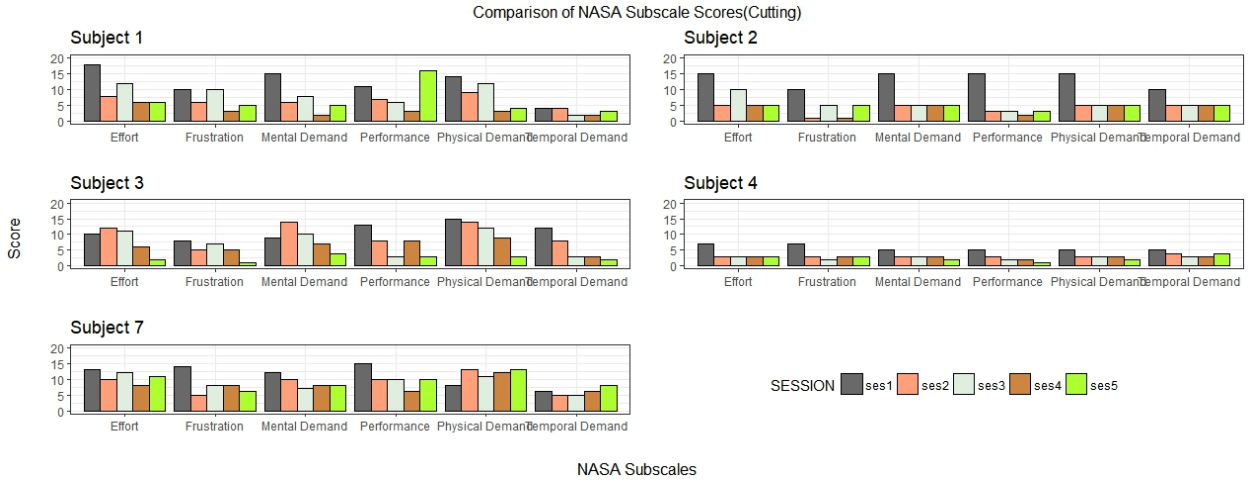


Figure 19: NASA TLX subscales comparison for different sessions for suturing task



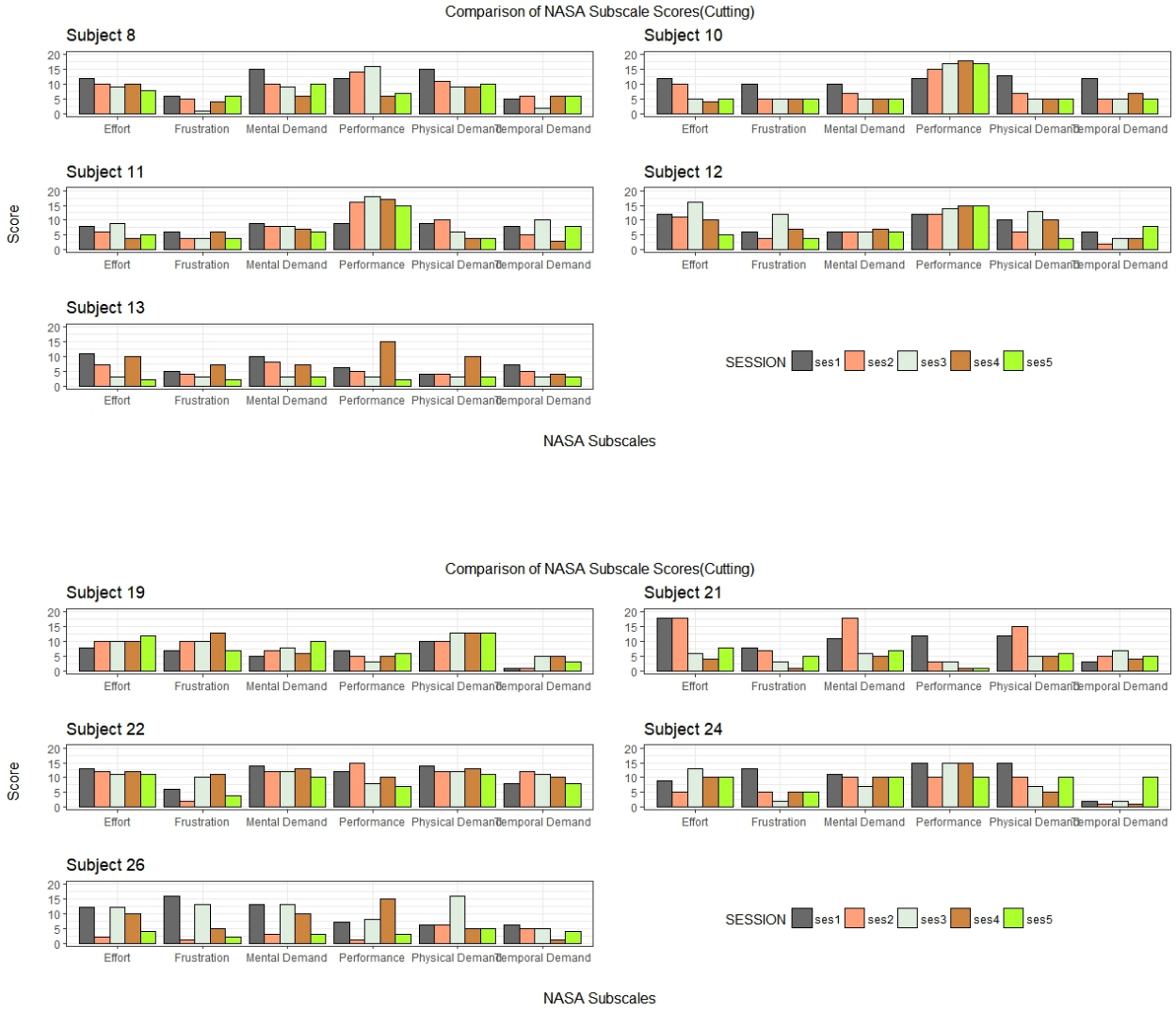


Figure 20: NASA TLX subscales comparison for different sessions for cutting task

5.4 Perinasal Perspiration (Stress) Signal Data

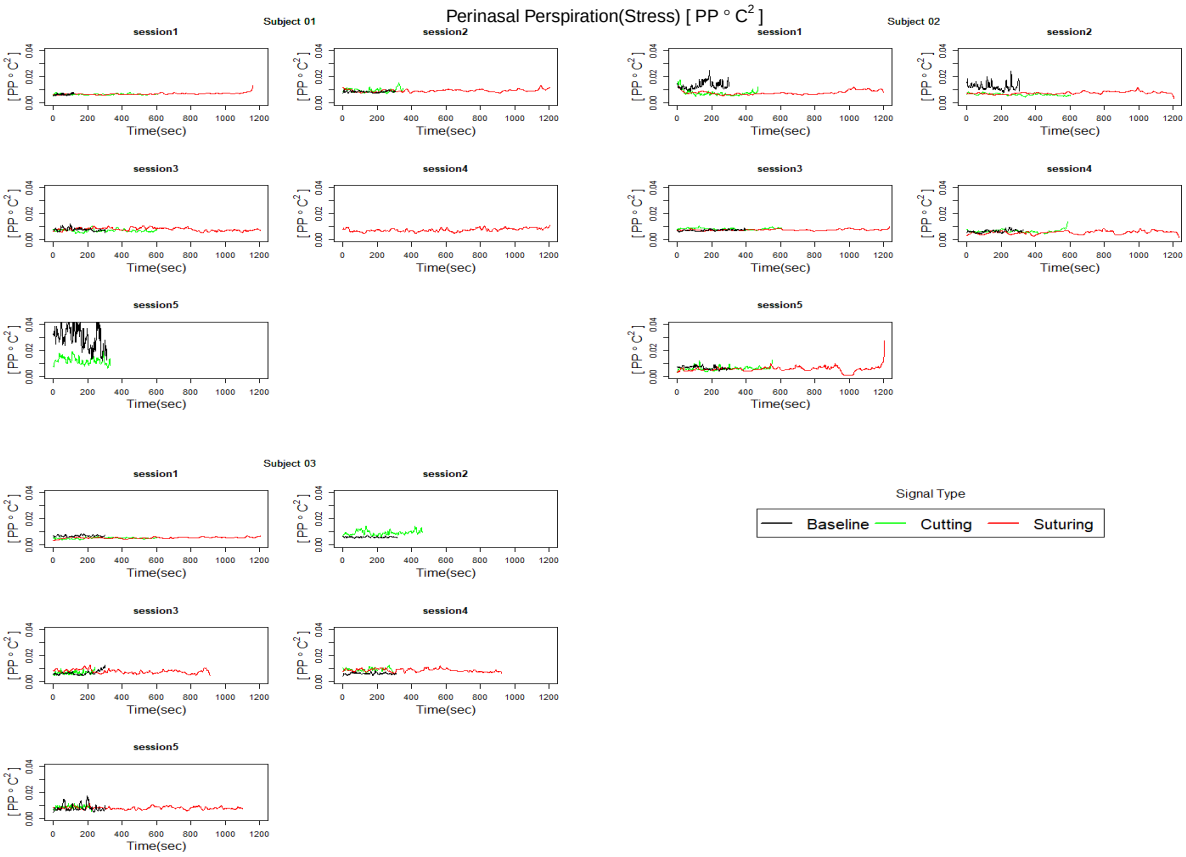


Figure 21: smoothed perspiration plots with time in seconds on x-axis

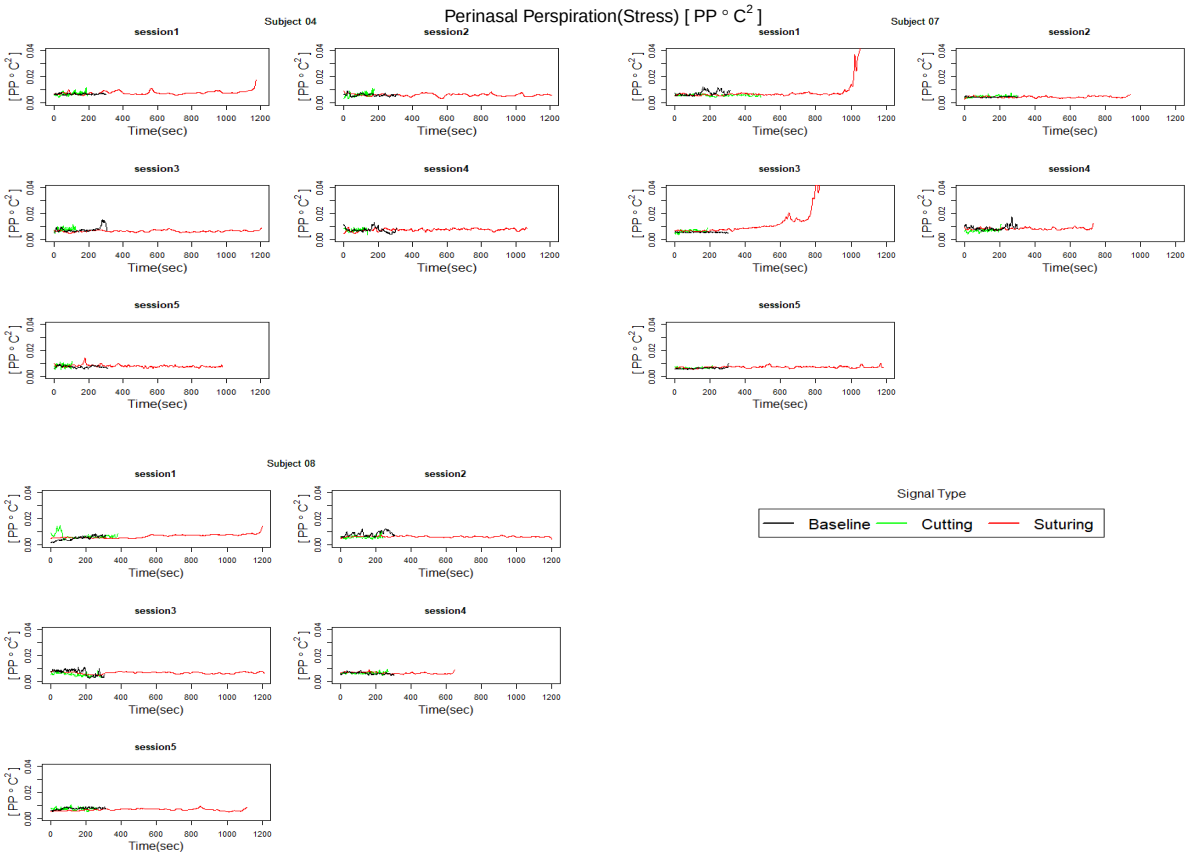


Figure 22: smoothed perspiration plots with time in seconds on x-axis

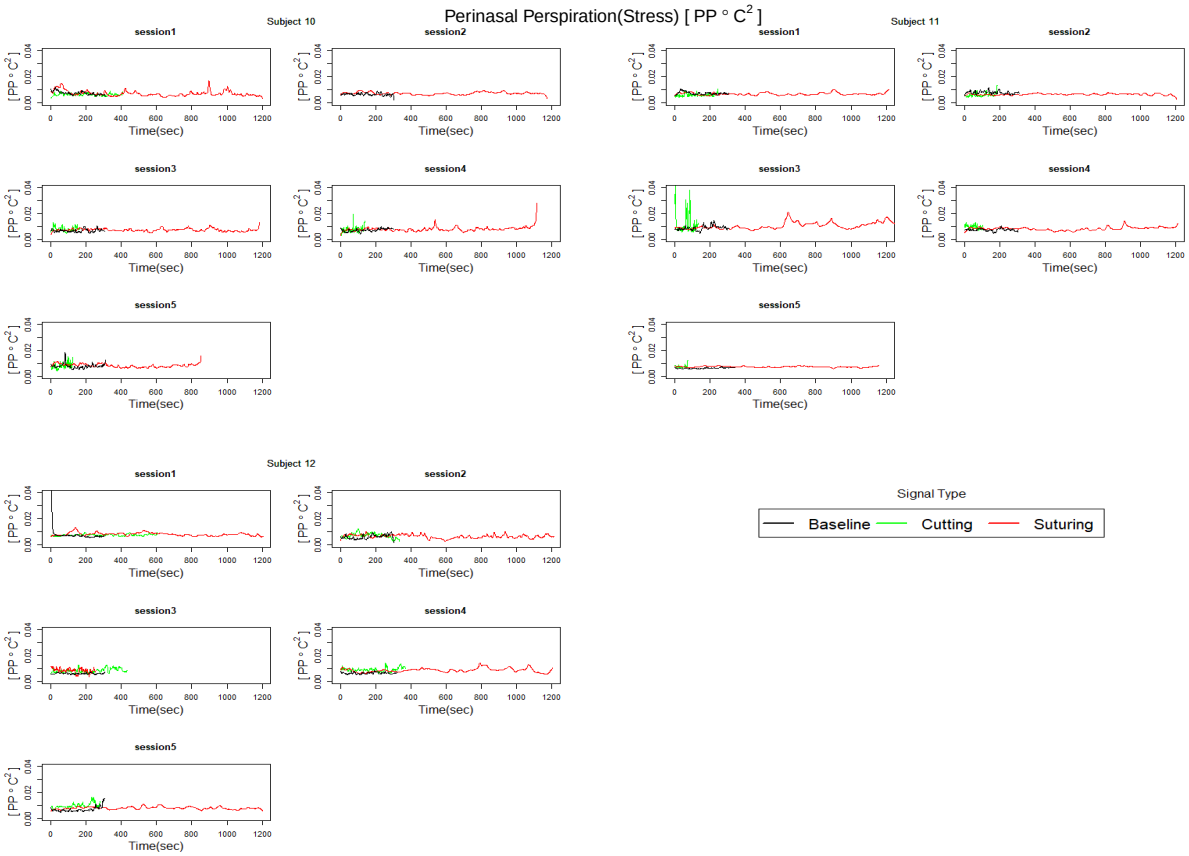


Figure 23: smoothed perspiration plots with time in seconds on x-axis

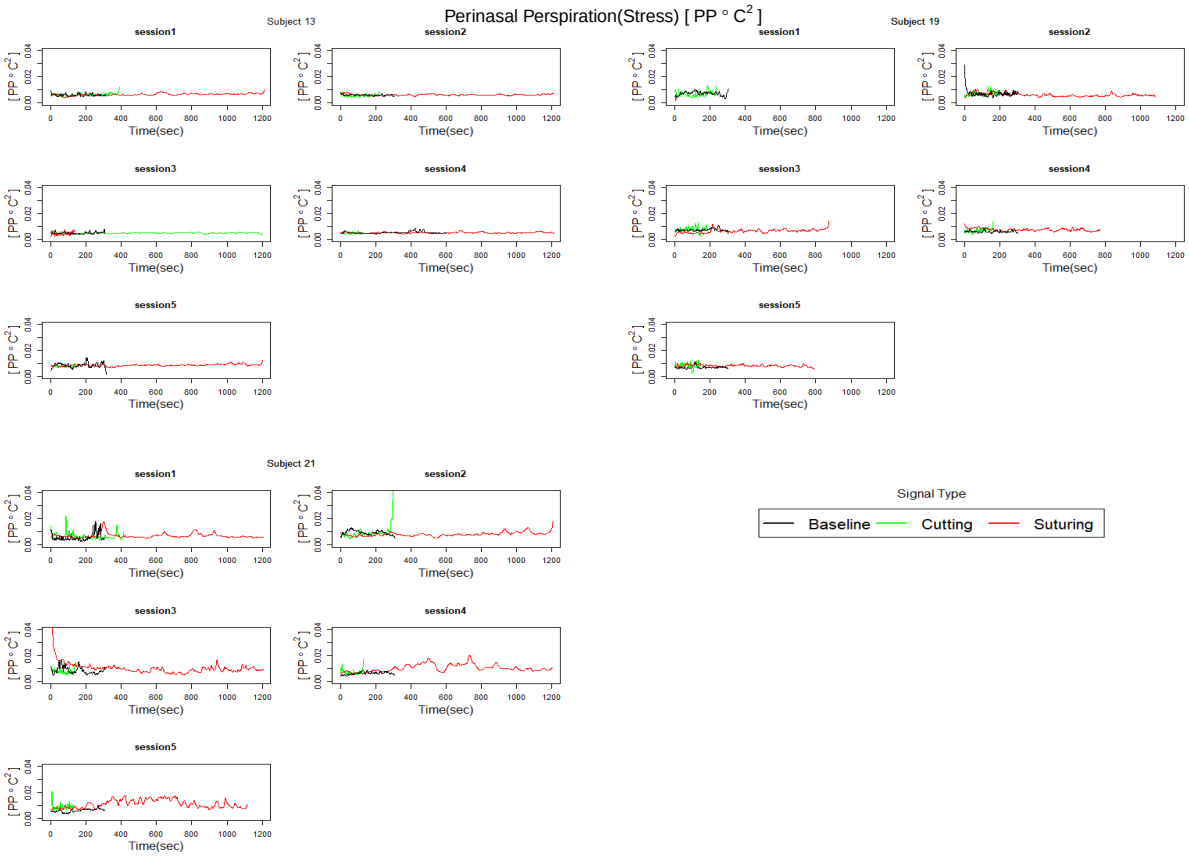


Figure 24: smoothed perspiration plots with time in seconds on x-axis

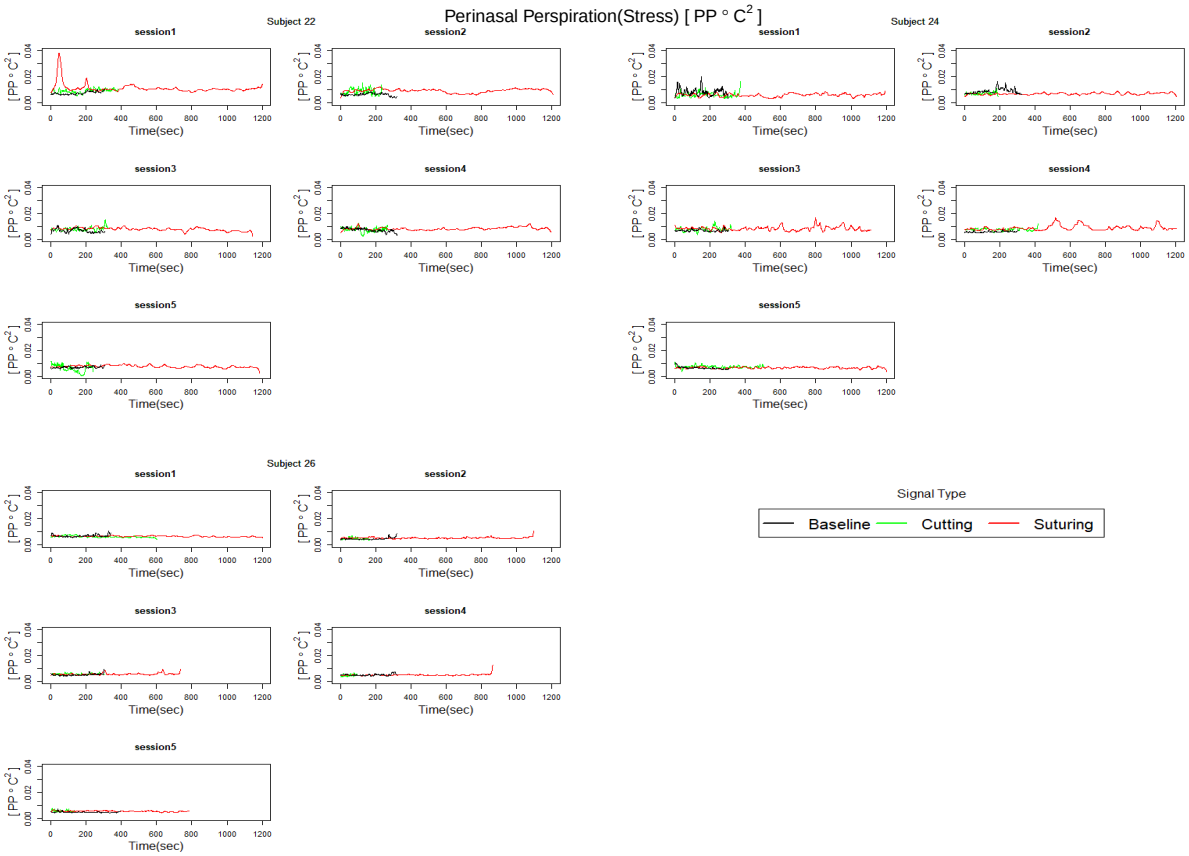
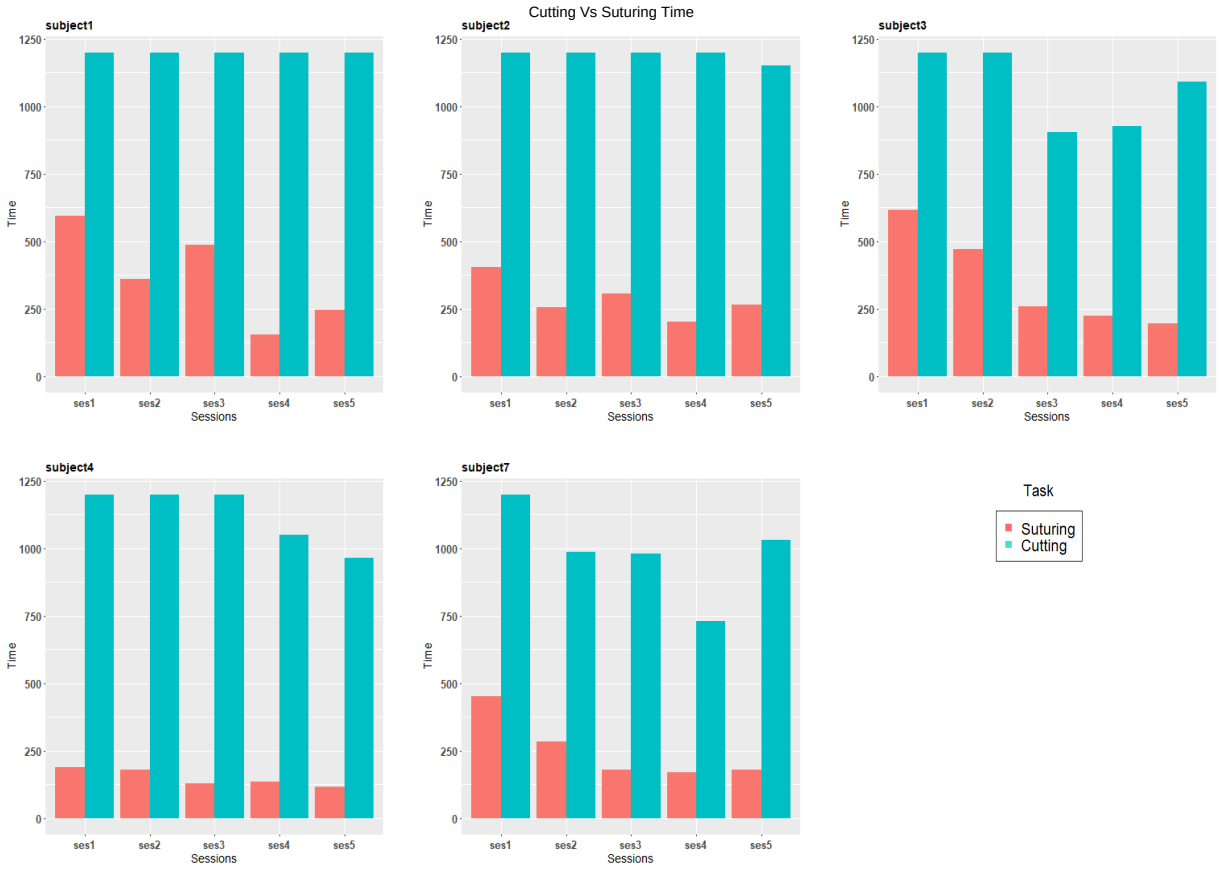
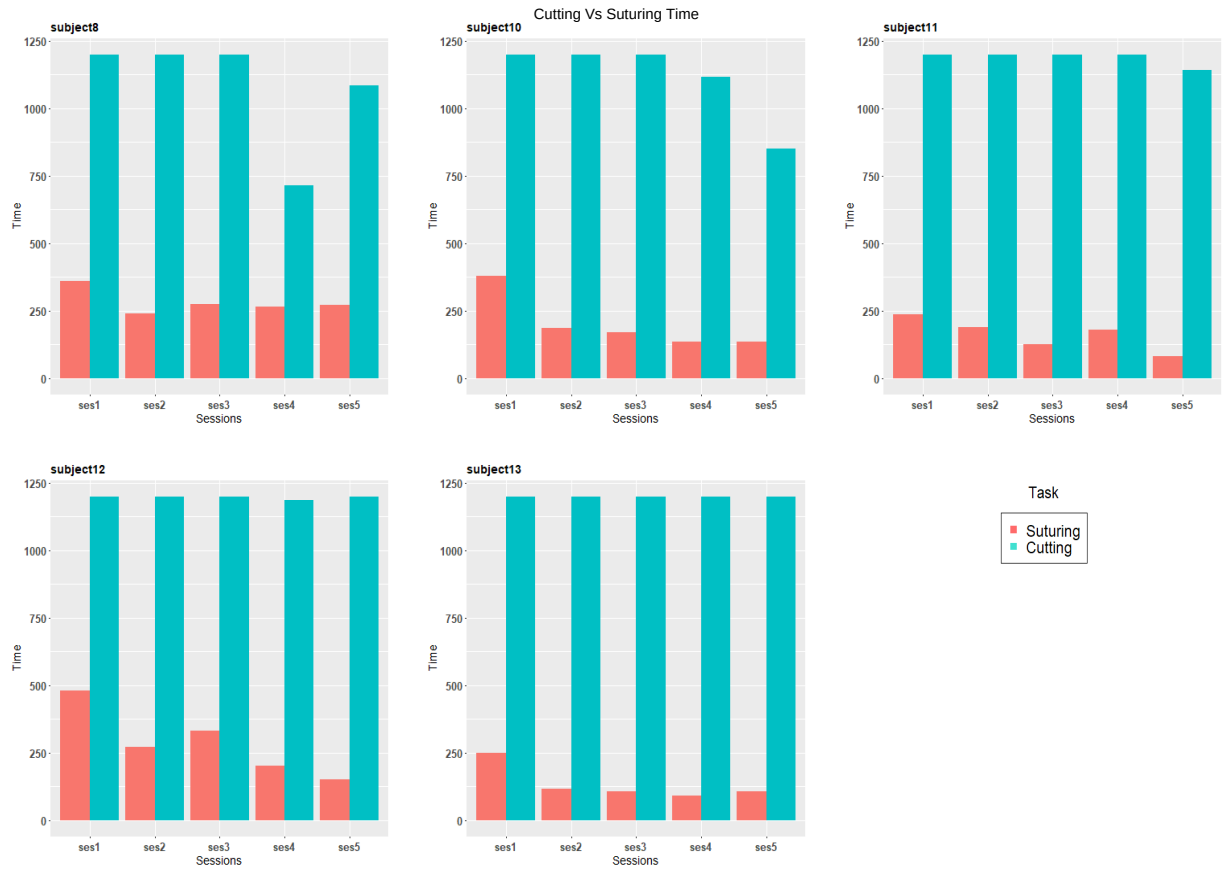


Figure 25: smoothed perspiration plots with time in seconds on x-axis

5.5 Performance Data

5.5.1 Cutting Vs Suturing Time




Cutting Vs Suturing Time
subject8
subject10
subject11
subject12
subject13
Time
Sessions
ses1
ses2
ses3
ses4
ses5
0
250
500
750
1000
1250
Task
Suturing
Cutting

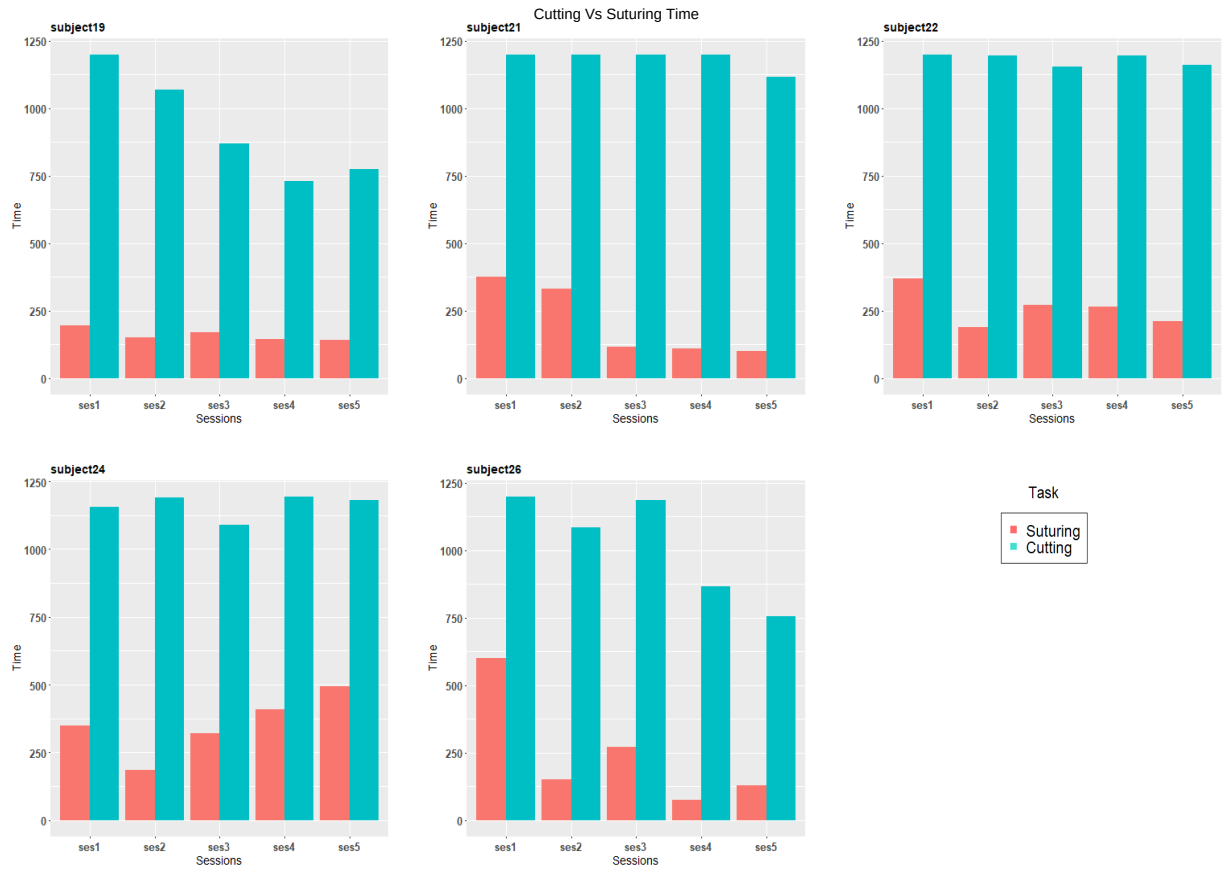


Figure 26: Plots depicting the time taken for the cutting and suturing tasks

5.5.2 Cutting Performance scores

Accuracy Plots for Scores (Cutting vs Suturing)

Figure 27: Plots depicting the performance scores by scorer1 and scorer2 for the cutting and suturing tasks

5.5.3 Accuracy plots for Number of sutures

Figure 28: Plots depicting the number of sutures performed by the subject in different sessions

The report is based on the 15 subjects who were involved in all the sessions of the task. There were total 22 subjects who participated in the research but were not included in the report since these subjects have not participated in all the sessions of the experiment and hence are considered as missing data. Since the data given has inconsistencies and missing data. A table to represent the inconsistencies is given below:

Subject	No. of Sessions	Session1	Session2	Seesion 3	Session 4	Session 5
1	5	✓	✓	✓	No baseline	No Suturing
2	5	✓	✓	✓	✓	✓
3	5	✓	No Suturing	✓	✓	✓
4	5	✓	✓	✓	✓	✓
5	3	✓	✓	✓	-	-
6	1	✓	-	-	-	-
7	5	✓	✓	✓	✓	✓
8	5	✓	✓	✓	✓	✓
9	2	✓	✓	-	-	-
10	5	✓	No cutting	✓	✓	✓
11	5	✓	✓	✓	✓	✓
12	5	✓	✓	✓	✓	✓
13	5	✓	✓	✓	✓	✓
19	5	Sut data	✓	✓	✓	✓
20	1	✓	-	-	-	-
21	5	✓	✓	✓	✓	✓
22	5	✓	✓	✓	✓	✓
23	4	✓	No Suturing	✓	NONE	-
24	5	✓	✓	✓	✓	✓
25	2	✓	✓	-	-	-
26	5	✓	✓	✓	✓	✓