A Comparison of Various Metrics Across Two Brain Supplement Versions

Sophia Leung
Springboard 2018

Introduction:

Nootropics are dietary supplements that are used for improving cognitive function and mental nutrition. In 2015, the international sales of these brain supplements are well over $1 billion USD and have been growing. These type of supplements are widely used by college students or anyone who wishes to increase their productivity. My mentor is currently working on a product whose company will remain anonymous. The brand aims to improve focus, energy, mood, and creativity. There are two trial versions of the supplement: 99 and 51. Trial 99 is the older version and Trial 51 is the newer version with some tweaked components. The supplement was given to users over a 21 day period and users answered survey questions at 1-week benchmark. The survey contained questions on how users would rate each given metric on an ordinal scale. The aim of this capstone project is to determine which version of the drug is better by comparing the metrics and quantify the correlation between the metric with the user's overall experience by using the overall experience score as a metric for success. With that in mind, Trial 51 is the newer version is the newer version of the supplement and will be assumed to be the better version under the null hypothesis. An analysis of the difference in the distribution of the metric's ratings will be able to determine which metrics are important or not.

Objective:

In order to determine which trial version is better the Overall Experience Rating will be used as the success metric:

- Each metric will be cross-compared to the overall experience rating, success will be determined if there is a correlation or likelihood of a user by its given metric will give a higher overall experience rating.
- For each metric, if the proportion of users in one trial is higher, then the metric is considered better for that respected trial.

The Dataset:

The Overall Experience Rating is the dependent variable where the submetrics will be used to predict trial version success. The success metric is on an ordinal scale of "Excellent, Good, No Effect, Not good, Awful". This rating scale was later converted into a binary scale in order to make the correlation analysis between the submetrics simpler. The 14 submetrics are the independent variables that will be used to compare across versions.

Data Cleaning:

The two datasets contained columns that were too long with words and were renamed to be more concise with the submetrics but also shorter. The "Version Duration" column contained categorical data that were in the format of a sentence where for example the value of, " 1 week (5 days)" was mapped to "1 week" in order to be consistent with the other values of 1 day, 2 weeks, and 3 weeks. The "Estimated Effect Duration" also contained categorical data that were again too lengthy. Instead, the values were changed in the format of numeric bins from "4 to 8 hours" to "4-8" such that an upper and lower numeric column was created for each bin for later analysis. In the "Current User Comparison" column, there are 4 possible values in which users were asked to rate their experience with the supplement if they have used the nootropic previously or as a non-user. In order to make the values more efficient, a separate column was created to categorize the users by "non-user" or "Nootropic user".

Table 1. The total number of observations and unique users for each trial

	# of Observations	**Unique Users**
Trial 99	102	40
Trial 51	133	47

Missing Data:

Since the data collected was in the format of a survey where users were asked to rate their experience over a period of time, some of the data are missing throughout the version duration. For example, User15 in trial 99 did not respond multiple metric questions and only answered for Day 1 and Week 1 data collection times. Therefore, when analyzing missing data for each metric, once a missing value was found, the user that was associated with that missing value was queried in order to check if the user has responded to other data collection times. If the user answered at two other data collection periods with the same consistent answer, then the missing value for that metric was imputed with the mode. If the user did not answer more for than two data collection times then the value was left missing so that it does not affect the data analysis.

Statistical Methods:

The general approach to the analysis to determine which trial is better by comparing the metrics between both trials. A variety of statistical tests were used in order

to determine the significance and effect size of each respected metric. However, there were limitations to the tests because the data within the trials are longitudinal and dependent. During the study, the users were asked to rate each metric over a period of time where data were collected at Day 1, week 1, week 2, and week 3. When comparing between the two trials, each metric is compared against each other with the Mann-Whitney U test, Difference in Proportions, and a Linear-by-Linear Association (ordinal association).

In order to use the Mann Whitney U test to determine if there is a difference in the distribution or median, the ordinal values were converted into numeric values to perform a ranking. Subsequently, if the results were significant, the Rank Biserial Correlation, Odds Ratio, and Common Language Effect Size were computed in order to quantify the effect size between the two trials. The Linear-by-Linear Associations test assess the relationship between the binary overall experience rating and the ordinal responses to the metric (Table 2). The odds ratio was then calculated to determine which trial were more likely to have users report an improvement to that metric when comparing between trials. Also, within trials the odds of how likely users who reported an improvement to a specific metric would be likely to give a higher overall experience (Table 5). Therefore, the odds ratio was calculated to compare how users within each trial with their overall experience and between trials for how users were more likely to report a positive experience with a given metric.

To determine if there is a difference in proportions for each metric, the proportion of users were grouped by their rating to that metric when comparing trials. For all metrics, with the exception of the Overall Experience, were not regarded with the longitudinal nature of the data. For example, the counts for Sleep Quality were binned into “Excellent”, “Okay”, and “Poor” for both trials. The difference in proportions was then compared between the two trials and Cohen’s H was calculated in order to quantify the effect size of the difference in proportions. Tables 2 and 4 shows the transformation of the data and the computation.

Table 2. Linear-by-Linear association for the sleep quality and overall experience in Trial 99

Sleep Quality	**Excellent**	**OK**	**Poor**	**All**
Overall Experience Rating				
High	46	40	4	90
Low	1	9	0	10
All	47	49	4	100

Table 3. Combining the Ok/Poor ratings in Sleep Quality to calculate the Odds Ratio within Trial 99

Sleep_Quality	**Excellent**	**OK/Poor**
Overall Experience		
High	46	44
Low	1	9

$$\text{Odds Ratio} = \frac{44 * 1}{46 * 9} = 9.409$$

Table 4.. Sleep Quality between both Trials to calculate Odds Ratio

Version	**Version 51**	**Version 99**
Sleep Quality		
Excellent	47	47
OK	72	49
Poor	13	4

Table 5. Combining the OK and Poor categories as one to calculate the Odds Ratio for Sleep Quality

Version	**Version 51**	**Version 99**
Sleep Quality		
High	47	47
Low	85	53

$$\text{Odds Ratio} = \frac{47 * 85}{47 * 53} = 1.603$$

Results:

There was a significant difference in the memory, capacity for productivity, and decisiveness median rankings of the users between Trial 51 and Trial 99 ($P < 0.05$ for Mann-Whitney U). All other metrics showed no significant difference in the distribution of how users would rate a given metric (Table 6). It's worth noting the odds ratio when comparing between trials to determine which trial users were likely to have a positive rating in Overall Experience, Physical Energy, and Sleep Quality is more favored towards Trial 99. Both Sleep Quality and the Estimated Effect Duration have a negative Rank biserial correlation showing that that as the values in Trial 99 decreases the values in Trial 51 increases.

Table 6. Comparing metrics between trials to see which metrics are different in median rankings.

	Mann Whitney U	P-Value	CLES	Rank Biserial	Odds Ratio
Overall Experience	6308	2.39E-01	53.50%	7.00E-02	1.00
Sleep Quality	7526.5	6.72E-02	42.98%	-1.40E-01	1.60
Focus, Attention, Concentration	6003.5	1.94E-01	54.97%	9.94E-02	0.54
Memory	4860.5	3.99E-04	63.54%	2.71E-01	0.33
Drive, Passion, & Motivation	6101.5	2.68E-01	54.23%	8.47E-02	0.33
Physical Energy	6694	9.56E-01	49.79%	-4.20E-03	1.01
Capacity for Productivity	5454.5	2.20E-02	58.77%	1.75E-01	0.45
Decisiveness	5566.5	3.11E-02	58.25%	1.65E-01	0.50
Verbal Fluency & Word Recall	5672	1.18E-01	56.03%	1.21E-01	0.61
Creativity & Insight	6219.5	4.93E-01	52.63%	5.26E-02	0.85
Interpersonal Capability & Empathy	5879	1.54E-01	55.46%	1.09E-01	0.74
Emotional Stability	6521	7.76E-01	51.09%	2.18E-02	0.85
Euphoria, Joy & Happiness	6189	3.50E-01	53.58%	7.16E-02	0.93
Estimated Effect Duration	6567	5.30E-01	47.56%	-4.87E-02	

Overall Experience:

In order to compare the overall experience comparison between the two trials, the ordinal responses were converted into binary responses: “high” and “low. “Good” and “Excellent” responses were binned into “High”. A “Low” rating were binned by“Awful”, “No effect”, and “Not good”. The frequency of the binary responses are then used to calculate the proportion of users that would give a high overall experience rating. To observe how the difference in proportions trended over time, the difference in proportions were calculated at various survey collection times at day 1, week 1, week 2, and week 3. Because the separated responses between the two trials have low frequencies, the data was then simulated using the Randomization test, where the number of “high” and “low” responses are shuffled according to the trial sample size of that respected time. All survey collection times saw no significant difference in the proportion of users who rated a high overall

experience between both trials (Table 7). The general overall experience proportion of high ratings was also found to not be significant (Figure 2).

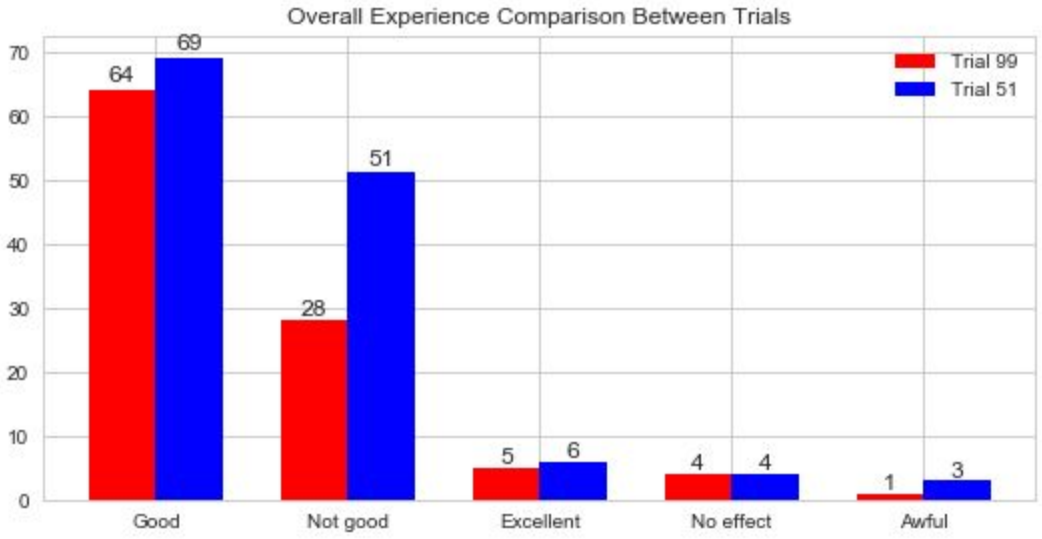

Figure 1. The count for each overall experience rating for both trials.

Under the null hypothesis the difference in proportions between the trials are equal and the alternative is the proportion of users in Trial 51 giving a higher overall experience rating is larger than Trial 99 (Table 7). At Day 1, the proportion of users that rated a higher overall experience in Trial 99 is higher than Trial 51 with a small effect size (Cohen's h=0.305). The simulated data for Day 1 has a p-value that is larger than alpha=0.05, therefore showing that the difference in proportion is relatively the same between the two trials. However, as the time approaches Week 1 and Week 2, Trial 51 has a higher proportion, but with an extremely negligible and small effect sizes as shown in the table. At week 3, there exist a small effect size (Cohen's h=0.355). In general, there is no statistical significance in the difference between the proportion of users that gave a high overall experience over time.

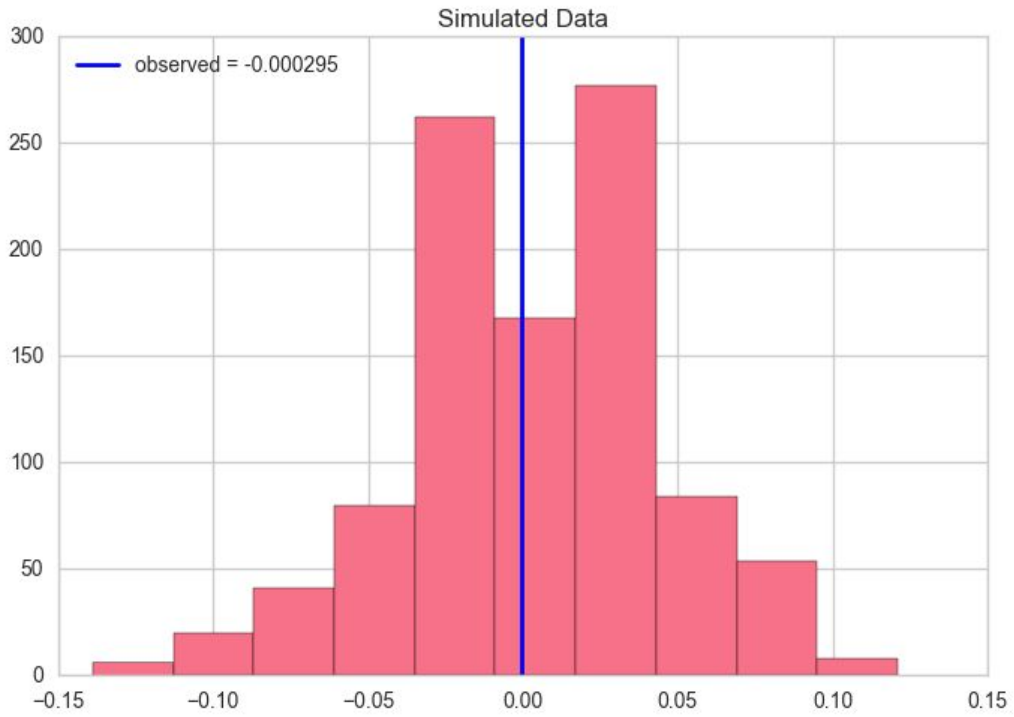

Figure 2. Simulated Data for by comparing the proportion of "high" ratings versus low between both trials.

Both the two proportion test and Mann Whitney U test shows there is no statistical significance between the two trials with p-values greater than 0.05 (Table 6). The Common Language Effect Size was also calculated between the two trials showing there is a 53.5% probability that Trial 51 has a larger proportion than Trial 99. But with such a low Cohen's h, this difference between the two trials is almost negligible.

In both trials, the number of users who rated their experience declined over time. Figure shows the number of users who rated an "Good" overall experience slowly declined as time approached week 3 in Trial 99 (Figure 3). However, all other ratings had a steady trend. In Figure, the number of "Good" ratings declined more rapidly at week 3 in Trial 51. The number of "Excellent" ratings increased at week 1 and later steadily declined at week 3. All other ratings were relatively steady with fewer than 5 rating counts (Figure 4).

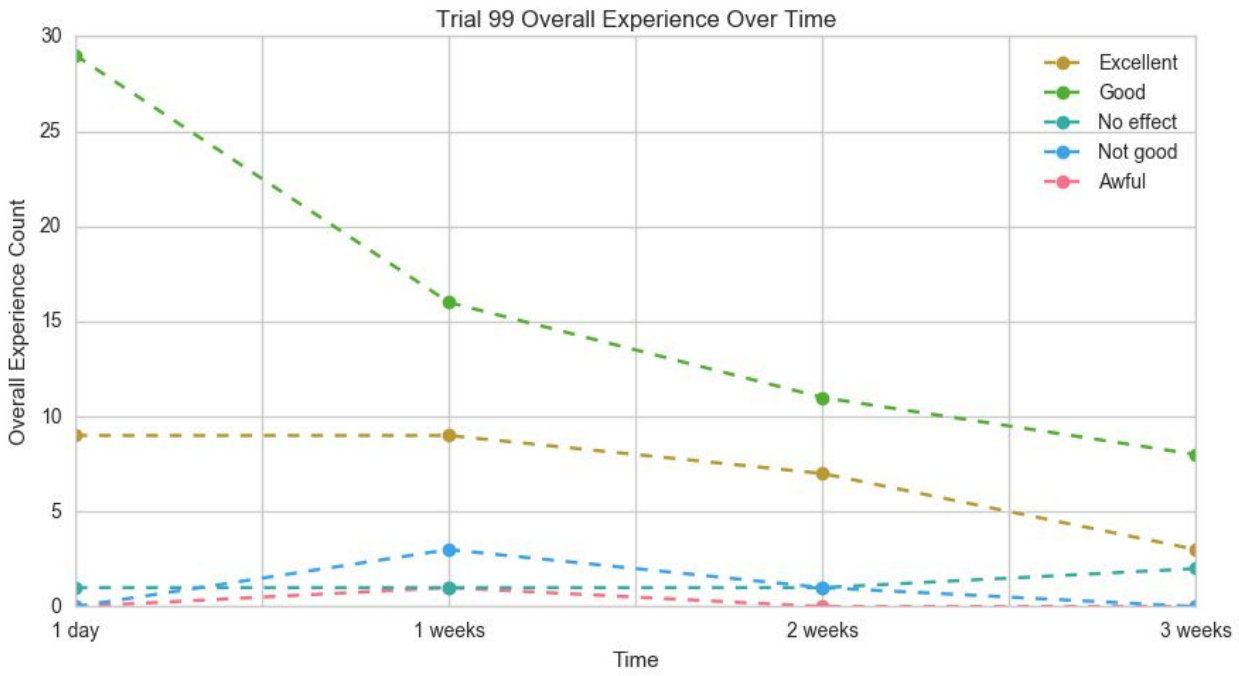

Figure 3. The count for the overall experience over time for Trial 99. Counts for positive ratings decreased over time, but an increase in no effect counts.

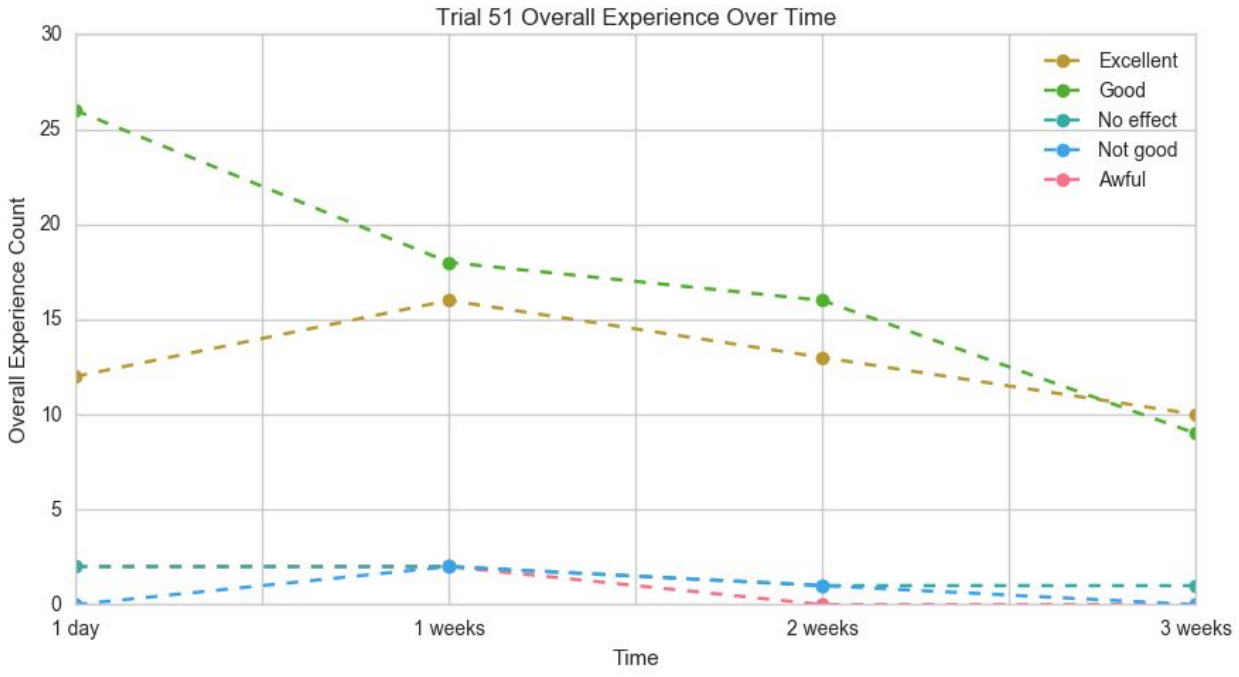

Figure 4. The overall experience rating count over time for Trial 51. Positive ratings counts decreased more than negative ratings.

Table 7. The difference in proportion for each version duration (data collection)

	Observed Difference	P-Value	Cohen's H
Day 1	0.06959	0.199	0.305
Week 1	-0.0166	0.70099	0.0456
Week 2	-0.03548	0.83799	0.01083
Week 3	-0.10384	0.95199	0.355

Sleep Quality:

There is no difference between the sleep quality median ranking between the two trials. Both trials have an "OK" median response. A negative rank biserial value of -0.140 shows there is a negative association between sleep quality and overall experience. In Trial 51, there exist an association between users who had excellent sleep quality and an higher overall experience rating; the odds ratio for that occuring is 7.56 (Table 8). The difference in proportions for the users with an excellent sleep quality is 11.3% higher in Trial 99. In contrast to the proportion of users who rated their sleep quality as poor is 5.8% higher in Trial 51.

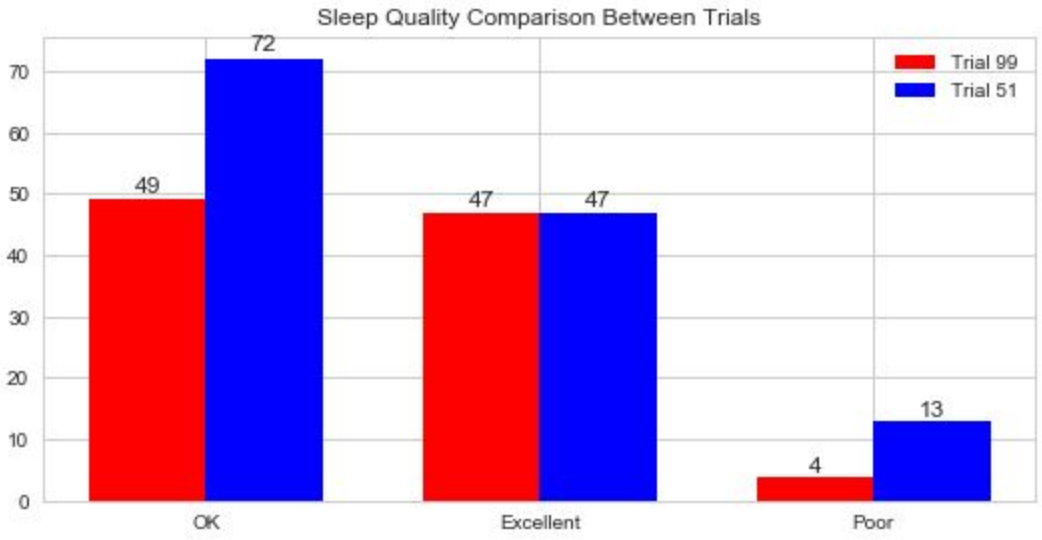

Figure 5. Sleep Quality rating count between trials. There are higher numbers of "OK" ratings.

Table 8. An Ordinal comparison and difference in proportions for Sleep Quality

Sleep Quality				
	Value	**P-Value**	**Odds**	**Cohen's H**
Linear-By-Linear Association				
Trial 99	762	0.89	9.41	-
Trial 51	16	3.03E-03	7.56	-
Prop. Difference				
Excellent	1.14E-01	4.00E-02	-	0.23
Okay	-5.55E-02	2.01E-01	-	0.11
Poor	-5.85E-02	4.52E-02	-	0.24

Figure 6. Sleep Quality rating counts over time where there is a sharp decrease in Excellent and OK ratings in Trial 99.

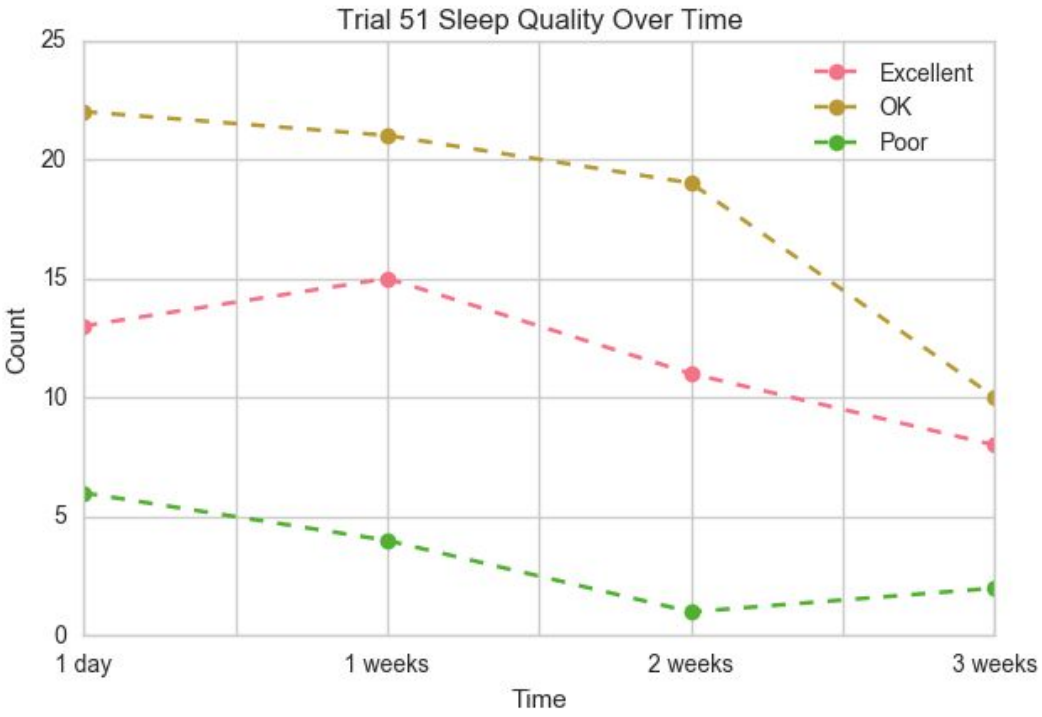

Figure 7. Sleep Quality rating counts over time in Trial 51. There are higher numbers of OK ratings that decrease steadily over time, but sharply as it approaches week 3.

Memory:

The proportion of users in Trial 51 who saw an improvement in their memory is 26.3% more compared to Trial 99. In contrast, the proportion of users who saw no effect on their memory is 26.7% higher in Trial 99 relative to Trial 51. However, there is no significant difference between the proportion of users in both trials who rated their memory as less good.

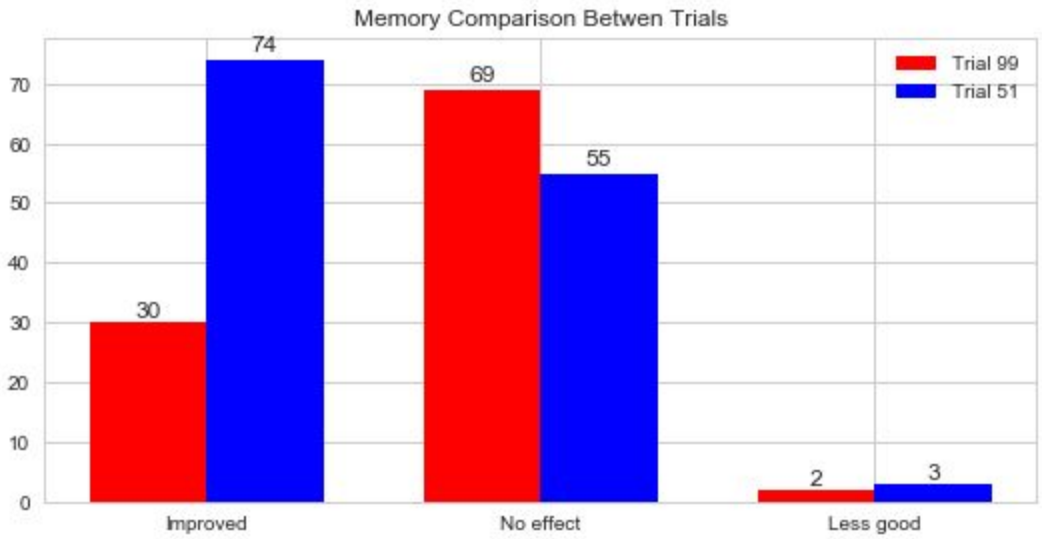

Figure 8. The counts for the memory ratings between both trials where most users found no effect for both trials on average.

Users who saw an improvement in their memory were more 19.04 times likely to give a higher overall experience rating in Trial 51. When comparing how users rated their memory between two trials, users in Trial 51 were 33.4 times more likely to give a higher rating than the users in Trial 99.

Over time, user ratings in Trial 51 increased sharply whereas Trial 99 had a steady trend. This trend though may have been affected by the number of users who continued to with the survey past week 2. As time approached week 3, the frequency of users who answered the survey declines.

Table 9. The Ordinal association and difference in proportions for Memory.

Memory				
	Value	**P-Value**	**Odds Ratio**	**Cohen's H**
Linear-By-Linear Association				
Trial 99	19.5	2.08E-01	4.21	-
Trial 51	22	1.22E-03	19.04	-
Prop. Difference				
Improved	-2.64E-01	3.03E-05	-	5.40E-01
No Effect	2.67E-01	2.67E-05	-	5.43E-01
Less Good	-2.93E-03	4.39E-01	-	2.03E-02

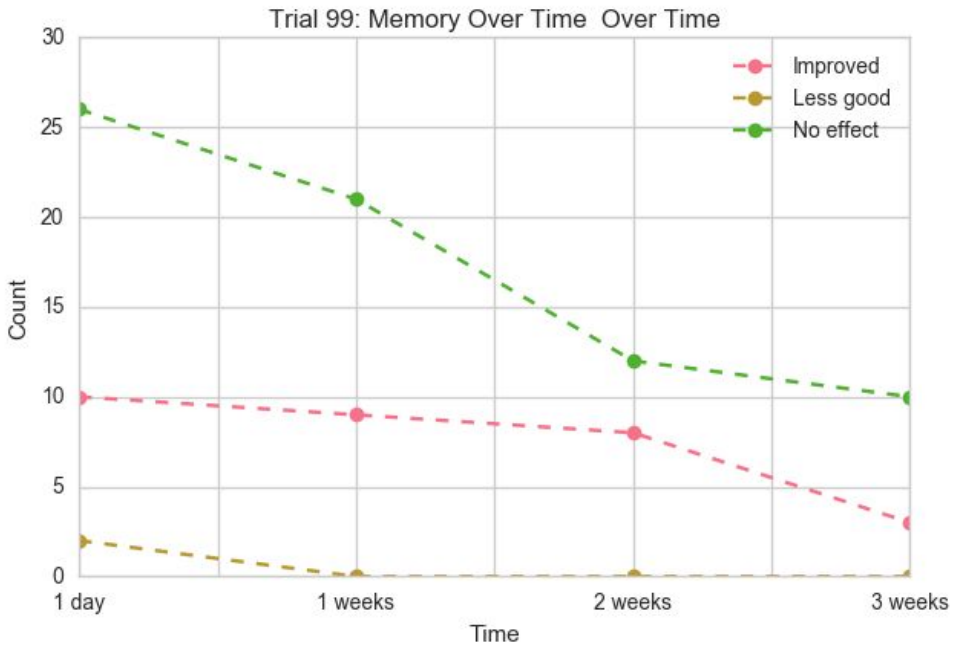

Figure 9. Memory rating counts over time where the number of No effect ratings decreases sharply as it approaches week 2 for Trial 99.

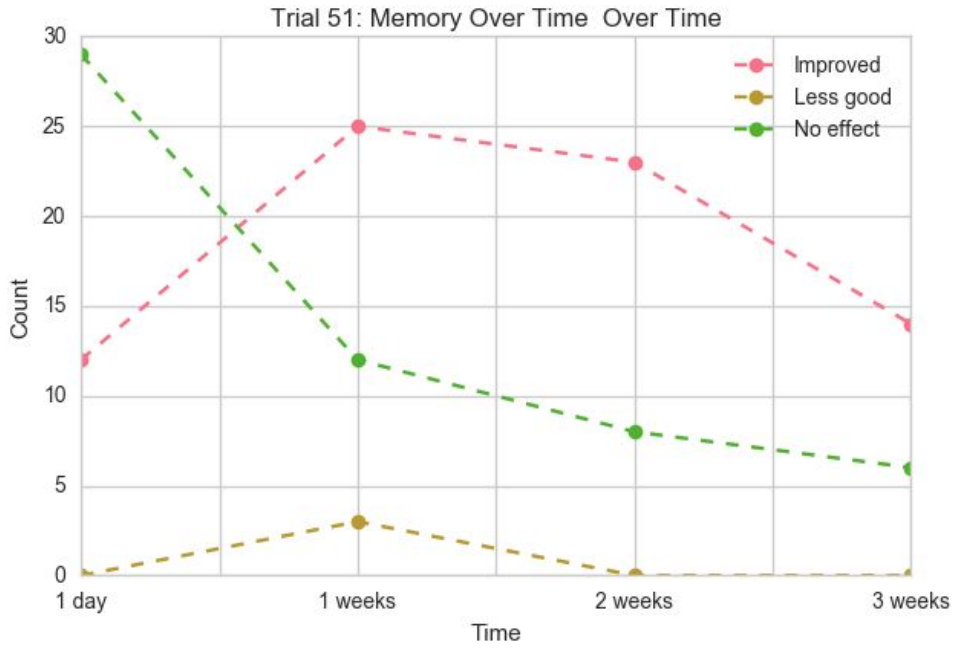

Figure 10. The memory rating counts over time where users who found an improvement increased at week 1 and hen sharply decreases at week 3 for Trial 51.

Capacity for Productivity:

The proportion of users in Trial 51 who saw an improvement in their capacity for productivity is 16.2% higher than Trial 99. Comparatively, there is no significant difference in the users who saw no effect on their productivity or as less good between the two trials. In both trials, there exist an ordinal association between users who saw an improvement in their productivity and giving an higher overall experience rating. Trial 51 users who saw an improvement were 76.5 times more likely to give an higher overall experience. Unfortunately, the odds ratio for Trial 99 could not be computed because there were zero counts for the low ratings in overall experience. When comparing between the trials, users in Trial 51 were 45 times more likely to see an improvement in their productivity relative to Trial 99.

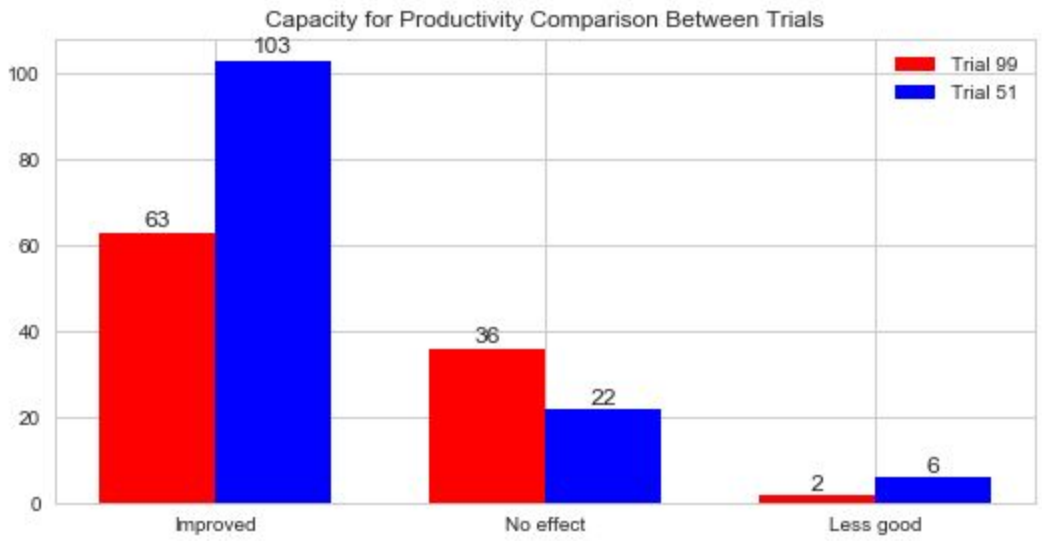

Figure 11. Capacity for Productivity count comparison between trials. Trial 51 has a higher proportion of improved productivity users than Trial 91.

Table 10. Ordinal Association and Difference in proportions for Capacity for Productivity.

Capacity for Productivity				
	Value	**P-Value**	**Odds**	**Cohen's H**
Linear-By-Linear Association				
Trial 99	19.5	3.23E-04	-	-
Trial 51	19.5	5.81E-07	76.5	-
Prop. Difference				
Improved	-1.62E-01	3.26E-03	-	0.36
No Effect	-2.60E-02	1.41E-01	-	0.15
Less Good	-5.70E-02	1.60E-01	-	0.13

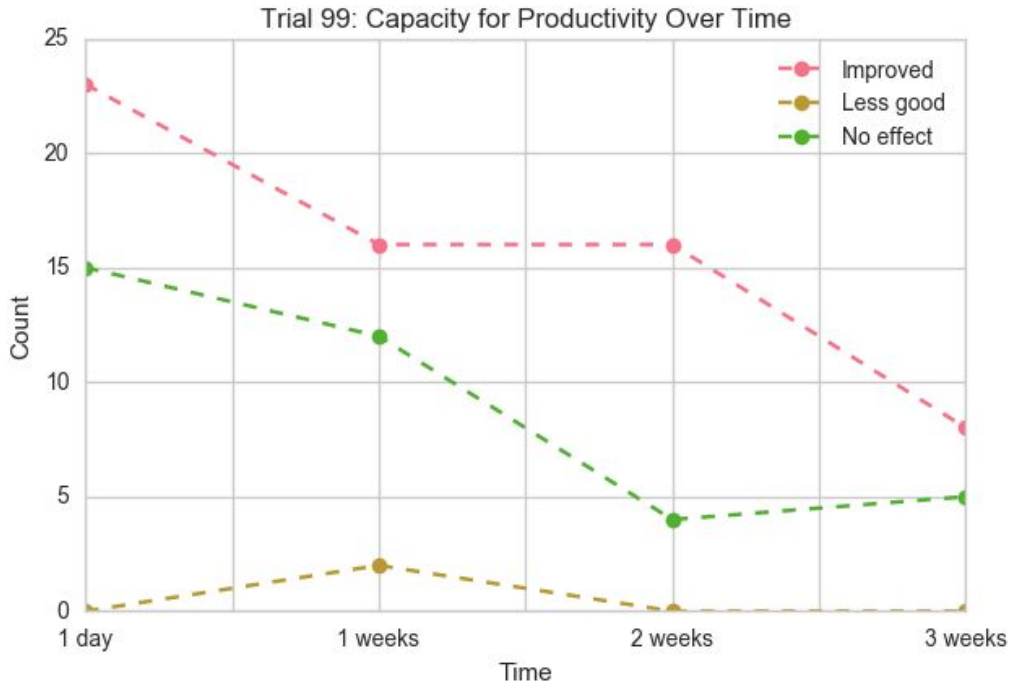

Figure 12. Capacity for Productivity over time where there is a sharp decrease at week 1 and 2 for Trial 99.

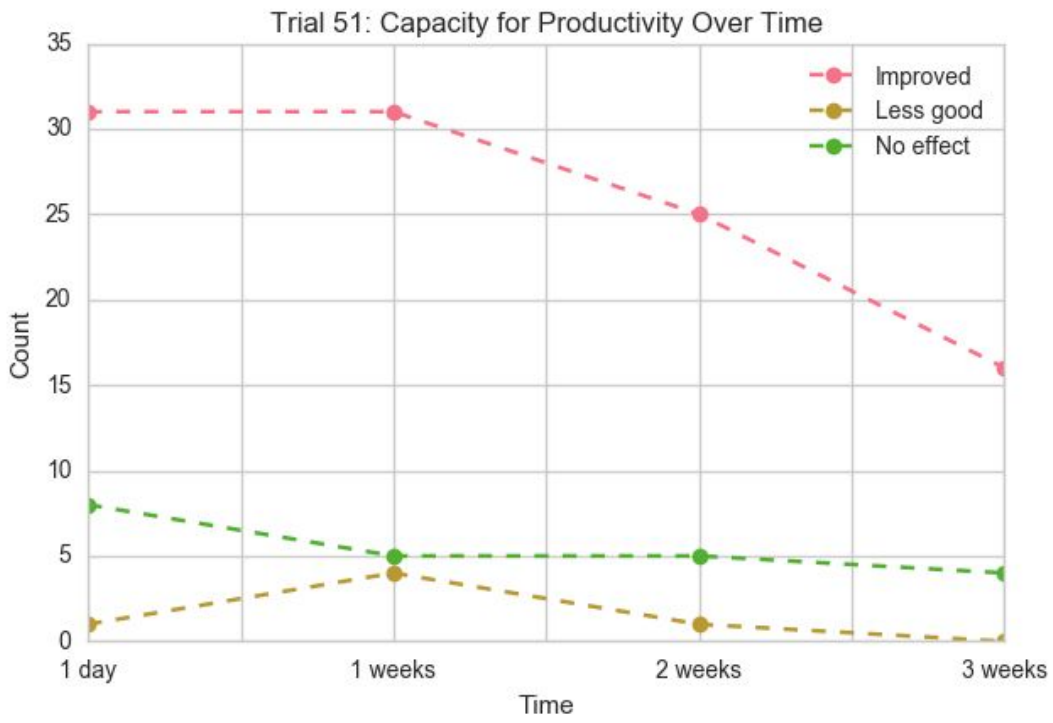

Figure 12. Capacity for Productivity over time where there is a sharp decrease at week 1 and 2 for Trial 99. .

Decisiveness:

There is a significant difference in proportions of the users who found an improvement or no effect on their decisiveness. The proportion of users who found an improvement is 15.7% higher in Trial 51 than Trial 99. On the other hand, the proportion of users who found no effect on their decisiveness is 17% higher in Trial 99. And lastly, there is no statistical significance in the difference in proportions for users who rated their decisiveness as less good.

Users in Trial 51 were 50 times more likely to see an improvement in their decisiveness compared to Trial 99. In addition, there exist an correlation between the users who gave a higher overall experience and seeing an improvement in their decisiveness. In Trial 99, users who saw an improvement were 13.75 times more likely to give a higher overall experience in contrast to Trial 51 which saw a higher odds ratio at 42.92 times more likely.

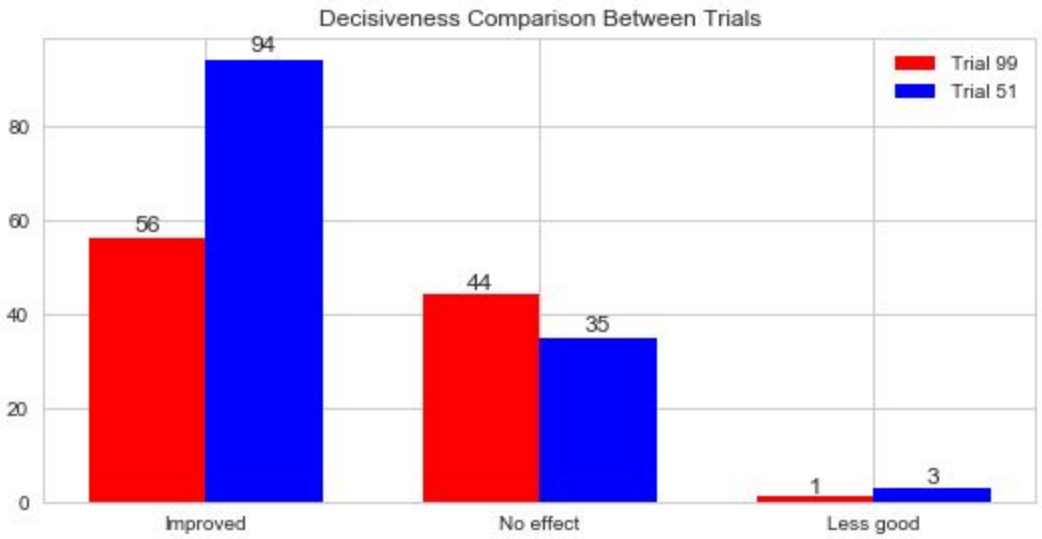

Figure 13. Decisiveness count comparison between trials. The number of improved users is extremely larger in Trial 51 than Trial 99.

Table 11. Ordinal Association and difference in proportions for Decisiveness.

Decisiveness				
	Value	**P-Value**	**Odds**	**Cohen's H**
Linear-By-Linear Association				
Trial 99	18.5	8.43E-03	13.75	-
Trial 51	22.5	6.88E-06	42.92	-
Prop. Difference				
Improved	-1.58E-01	6.38E-03	-	0.33
No Effect	1.70E-01	3.22E-03	-	0.36
Less Good	-1.28E-02	2.28E-01		0.10

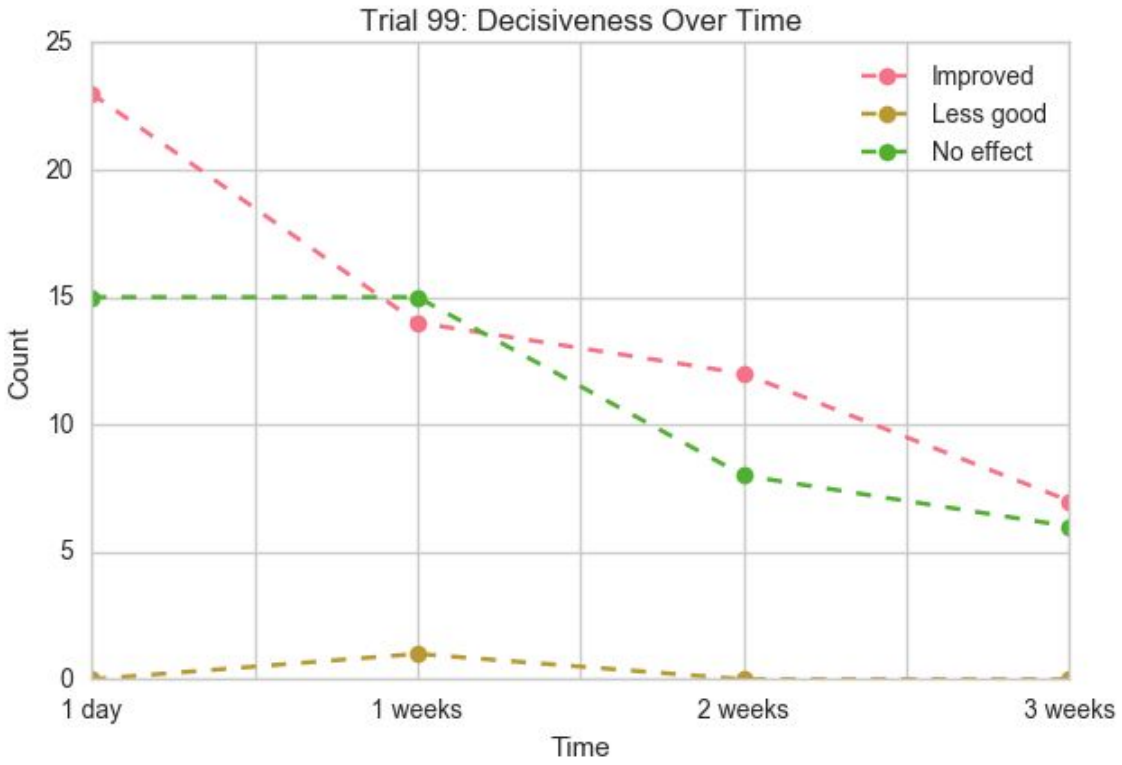

Figure 14. Decisiveness count ratings over time for Trial 99.

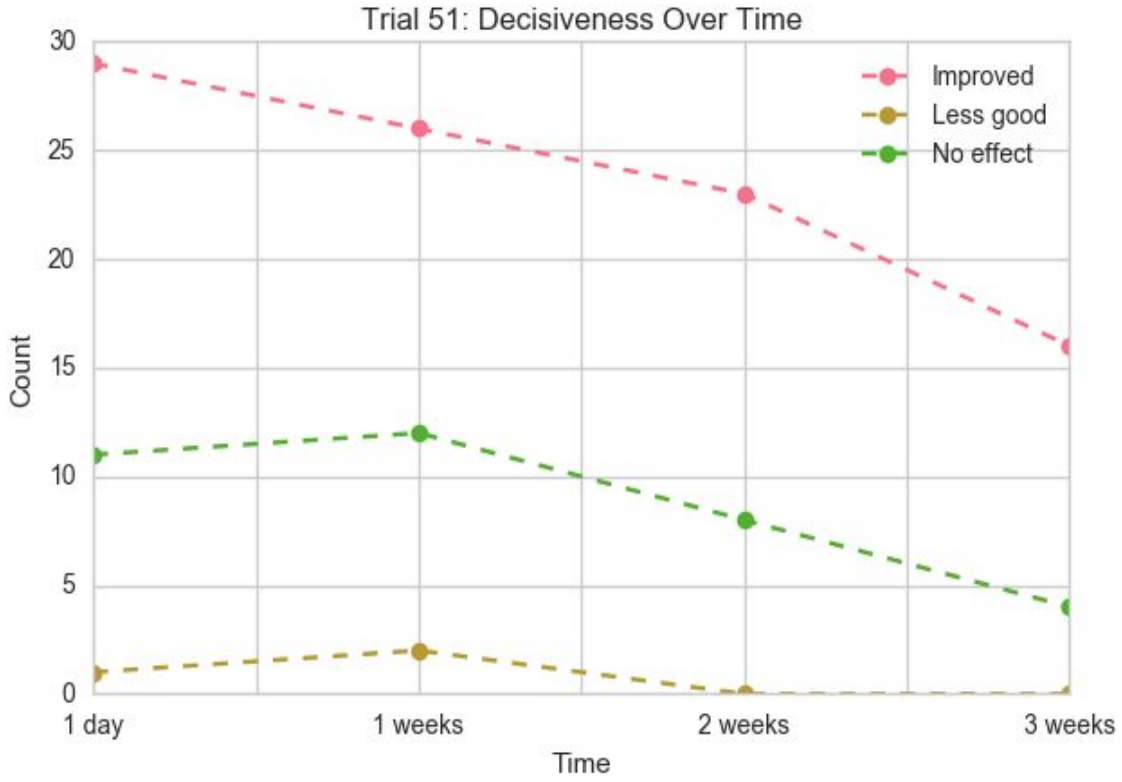

Figure 15. Decisiveness count ratings over time for Trial 951

Estimated Effect Duration:

In both trials, the median estimated effect duration is 4 to 8 hours. Trial 99 have a significant correlation between a higher overall experience and estimated effect duration whereas Trial 51 does not. There is no significant difference in the proportion of users for all estimated effect duration time ranges.

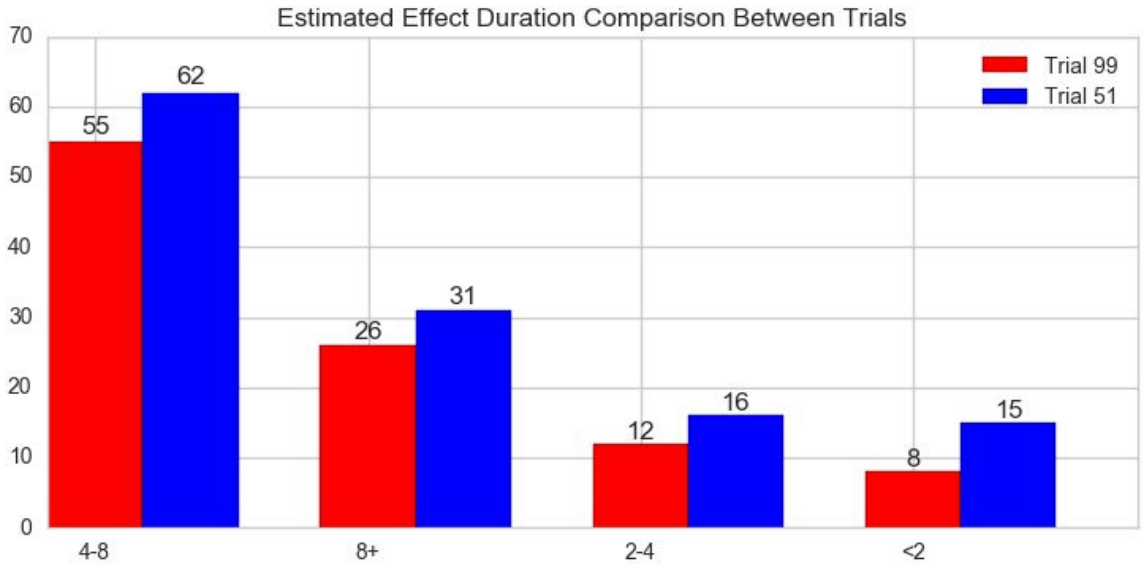

Figure 16. The Estimated effect duration counts between trials where on average is 4-8 hours.

Table 12. The Ordinal Association and difference in proportions for the estimated Effect duration.

Estimated Effect Duration				
	Value	**P-Value**	**Odds**	**Cohen's H**
Linear-By-Linear Association				
Trial 99	19	1.02E-02		-
Trial 51	17	2.07E-01		-
Prop. Difference				
<2 Hours	-4.18E-02	1.52E-01	-	0.14
2-4 Hours	-1.02E-02	4.09E-01	-	0.03
4-8 Hours	4.46E-02	2.53E-01		0.09
8+ hours	7.43E-03	4.49E-01		0.02

The odds ratio was not calculated for the linear-by-linear association for both trials.

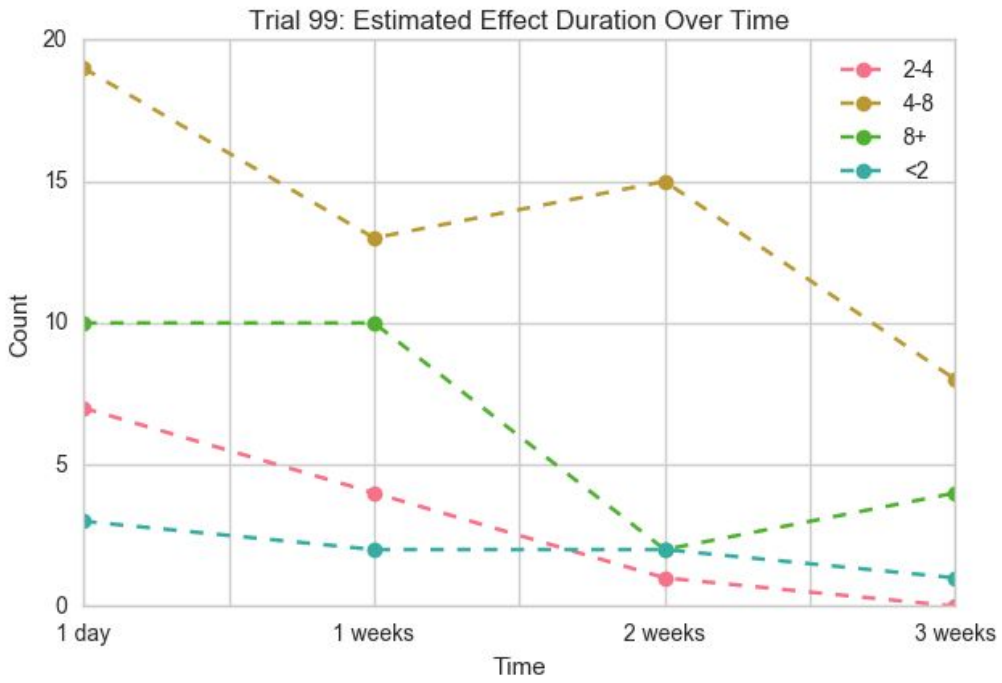

Figure 17. The Estimated effect duration counts over time, where there is a sharp decrease in 4-8 hours at week 2 for Trial 99.

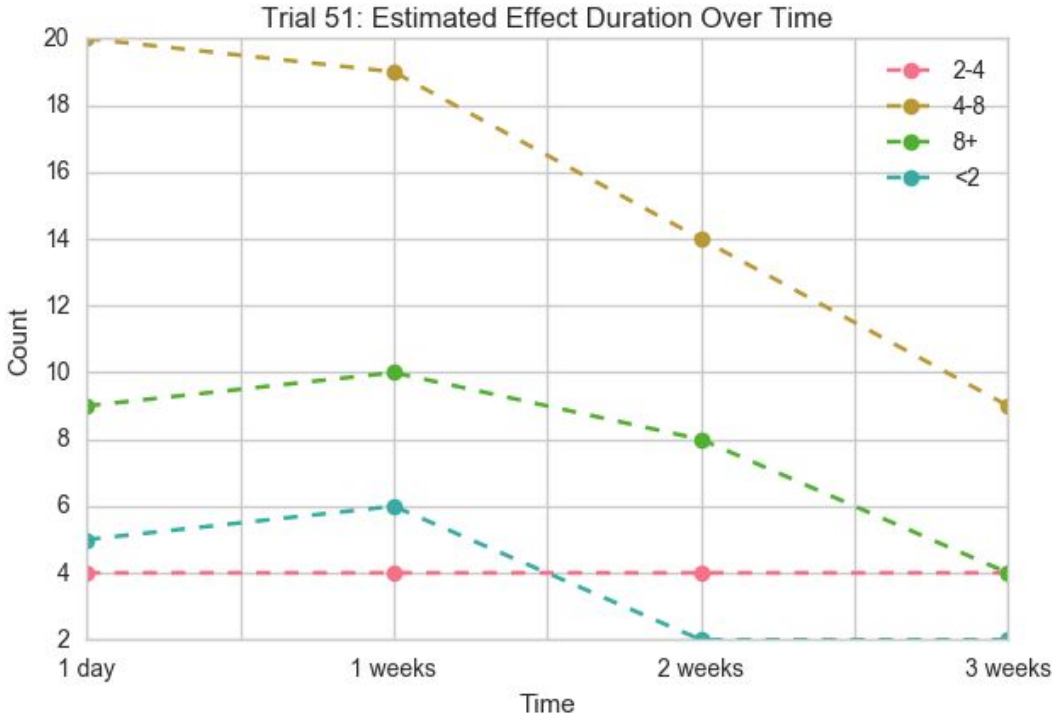

Figure 18. The Estimated effect duration counts over time, where there is a sharp decrease in 4-8 hours at week 1 for Trial 51.

Machine Learning:

The goal was classifying users by a high or low overall experience based on the submetrics by using SVM, Logistic Regression, and the Random Forest Classifier. Both trials were combined into one single DataFrame in order to collect more data. At first, only the data from Trial 99 was used, but it was quickly discovered that it was not enough data. Once the two DataFrames were combined, the ordinal values were converted into numeric values and ordered. Missing values were denoted as -1. Even when combining the two datasets, the classifier had a class imbalance issue where there were more "high" overall experience ratings. The minority "low" class in the training set was then up-sampled by use of the SMOTE algorithm which creates synthetic samples of the minority class.

Out of all the models, the SVM and Random Forest out-performs the other models by recall, accuracy, and precision. These two models for all scoring metrics seem to have training and test scores that generalize well. In contrast to the Logistic Regression model, which for all scoring metrics, shows that it has a tendency to overfit. This may be due to not standardizing the data. But the logic is that since the values are in ordinal, with only 3 possible answers, it should be the same in general for all features. Which is why it was decided not to because it appears the data is already standardized. Lastly, the Multiclass Random Forest model when measuring its accuracy shows that the model has a tendency to underfit. The model may not be able to handle the 5 ordinal scales for the overall experience and the 14 other features with 3 level ordinal ratings.

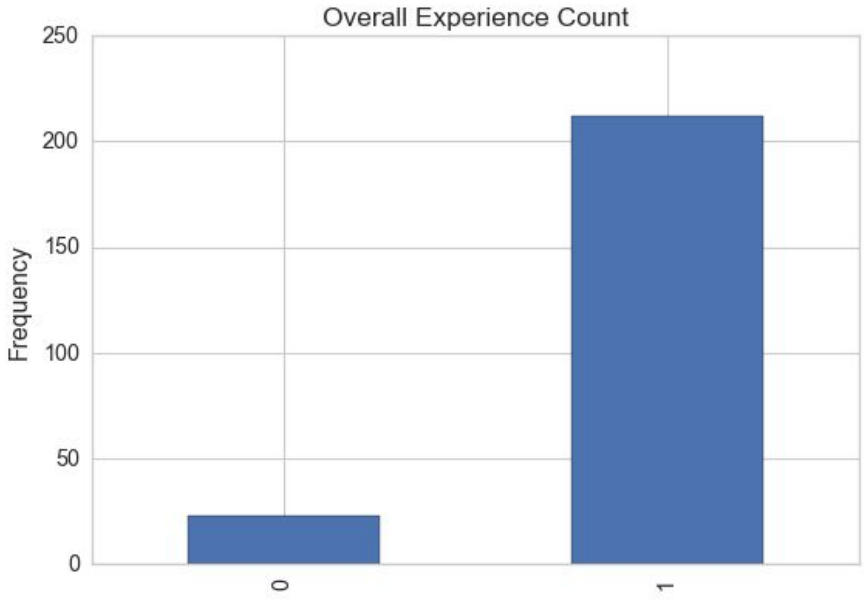

Figure 19. Count of the overall experience for both trials.

Table 13. The recall score metric for the classifiers for the training and test set

Recall	Training	Test
SVM	99.32%	95.77%
Logistic Regression	99.32%	88.73%
Random Forest	100.00%	90.14%

Table 14. The accuracy score metric for the classifiers for the training and test set

Accuracy	Training	Test
SVM	99.66%	94.37%
Logistic Regression	98.65%	84.51%
Random Forest	97.56%	94.37%
Random Forest (Multiclass)	63.41%	66.20%

Table 15. The precision score metric for the classifiers for the training and test set

Precision	Training	Test
SVM	100.00%	92.96%
Logistic Regression	100.00%	84.51%
Random Forest	97.52%	95.77%

Feature Selection:

Since there are 14 features that are used for the classifier, a wrapper method for feature selection was used in order to determine which are the most important based on the various scoring metrics. The Recursive Feature Elimination method finds the best subset of features that have the highest performance by ranking them. Scikit learn has an algorithm that combines both the RFE and cross-validation process to select the best number of features. In addition, the Random Forest and Extra Trees classifier includes a feature importance attribute.

In the SVM model, the recall scoring metric has the highest score with 4 as the optimal number of features. These features include: “Drive/Passion/Motivation, Focus/Attention/Concentration, Interpersonal Capability/Empathy, and Physical Energy (Figure 20).

In addition to Physical Energy and Interpersonal Capability/Empathy appears the most frequent for all scoring metrics in the SVM linear model. In the Random Forest classifier, Physical energy is the top ranking feature followed by Decisiveness and the Estimated Effect Duration. However, the recall scoring metric has the highest score with only 1 optimal number of features that is Physical energy. For the Extra-Trees classifier, recall is also the top scoring metric with 2 optimal number of features that include: "Focus/Attention/Concentration, and Physical Energy. The top feature with the importance ranking is the Estimated Effect Duration followed by Physical Energy, and Focus/Attention/Concentration. Lastly, the Logistic Regression model also has Recall with the highest score with 14 optimal number of features, but from the graph, the optimal number of features could be reduced to 4. In all the models, it appears that Physical energy, in general, is the most important feature.

After using the selected features for the SVM model, the training recall decreased from 99.32% to 96.58%. And the test recall from 95.77% to 97.18%. In the new model with the selected features, the SVM has a slight tendency to underfit. The Random Forest recall score did not change even after using the important features.

Table 16. The number of optimal features for each classifier and scoring metric.

Model	Accuracy	ROC AOC	F1	Precision	Recall
SVM-Linear	2	5	2	11	4
Logit	14	8	14	14	14
Random Forest	8	12	1	13	2
Extra Trees	14	2	14	7	2

Table 17. The recall score only using a subset of the data with important features.

Recall	Training	Test
SVM	96.58%	97.18%
Random Forest	100.00%	90.14%

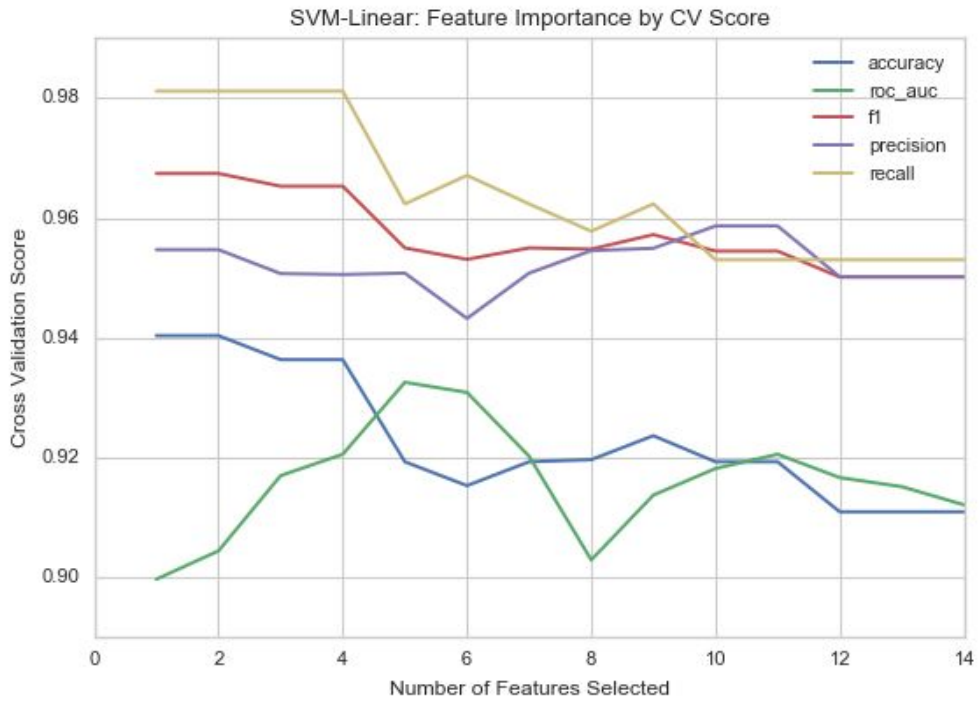

Figure 20. The optimal number of features for the SVM classifier for each score metric where the best score metric appears to be recall with 4 optimal features.

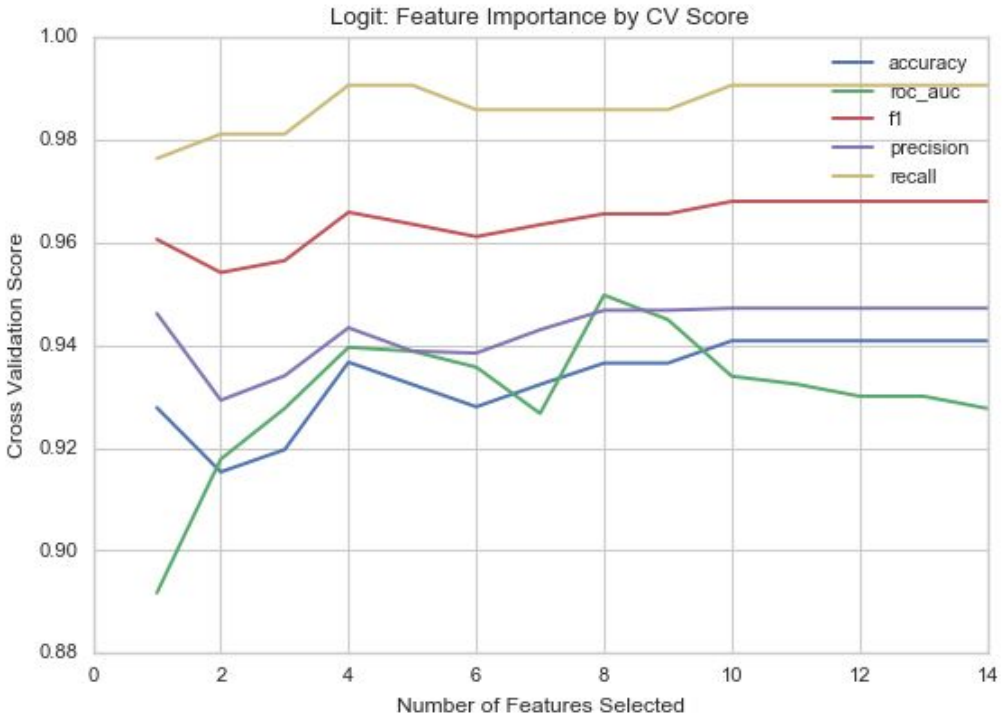

Figure 21. The optimal number of features for the Logistic Regression model for each score metric where the best score metric appears to be recall with 14 optimal features..

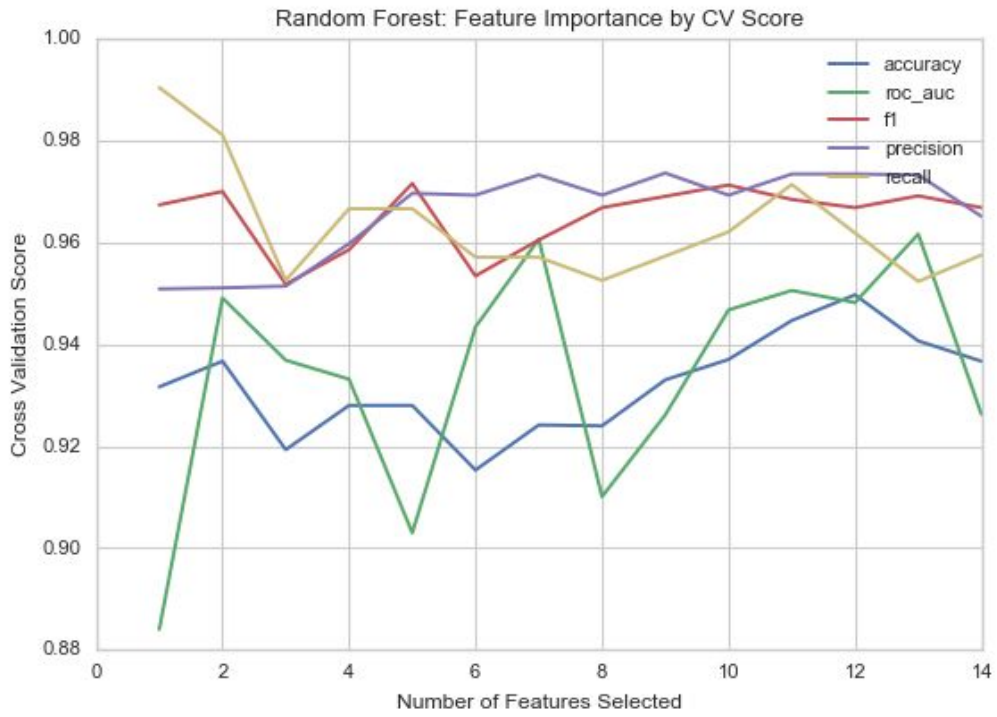

Figure 21. The optimal number of features for the Random Forest classifier for each score metric where the best score metric appears to be recall with 2 optimal number of features.

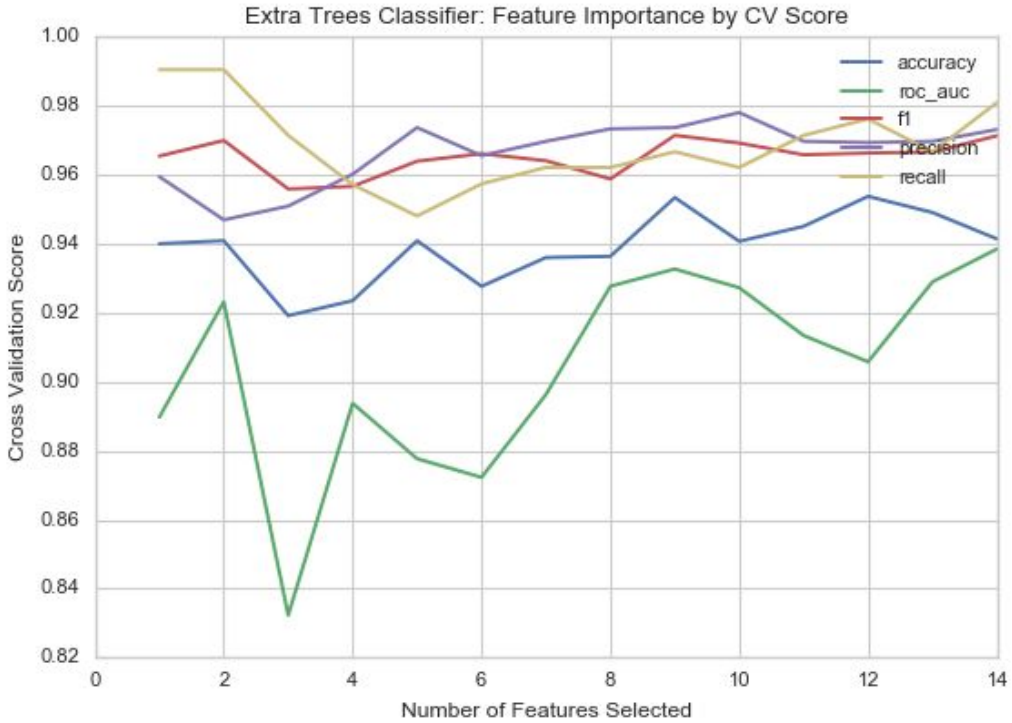

Figure 22. The optimal number of features for the Extra-Trees classifier for each score metric where the best score metric appears to be recall with 2 optimal number of features..

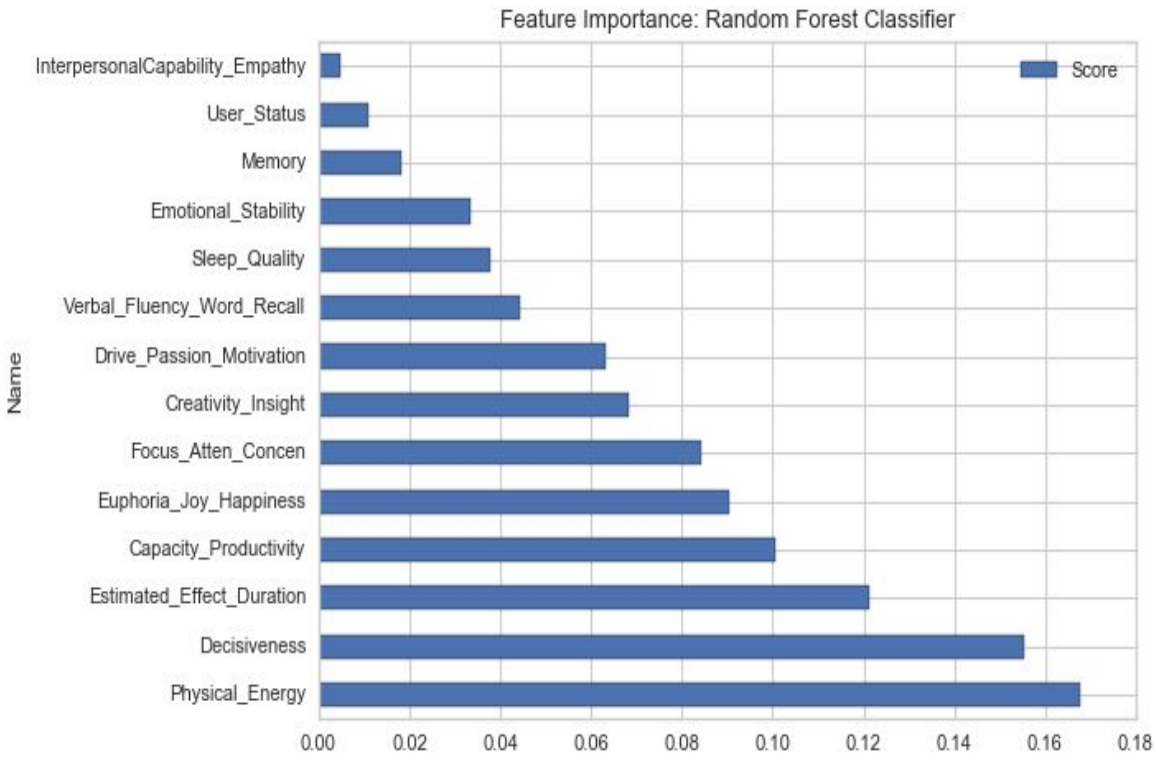

Figure 23. Feature importance for the Random Forest Classifier with Physical Energy and Decisiveness being the highest.

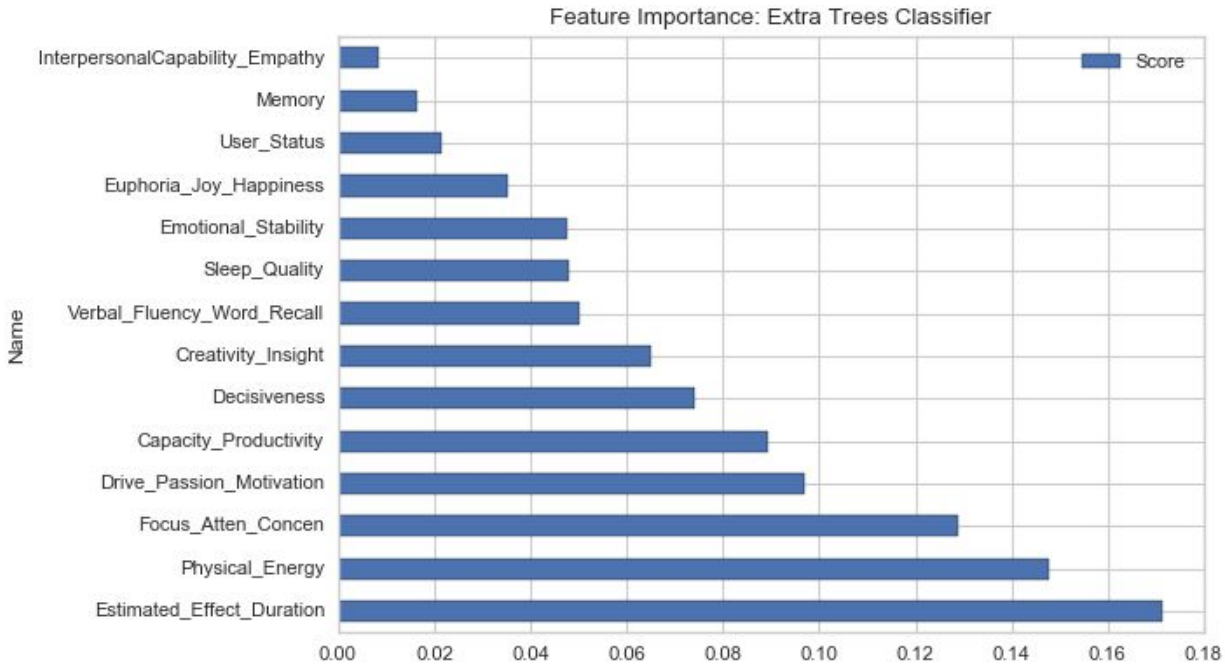

Figure 24. Feature importance for the Extra Trees Classifier with the Estimated Effect Duration and Physical Energy being the highest.

Conclusion:

There exist a difference in the submetrics: Memory, Decisiveness, Focus/Attention/Concentration, Verbal Recall, and the Capacity for Productivity between the two trials such that there are a higher number of users in Trial 51 who found an improvement in these metrics. For the ingredients in the supplement that causes an increase in these metrics are stronger in Trial 51 and should be kept for all future versions. However, the Sleep Quality and Physical Energy, albeit have no difference in the median rankings, has an odds ratio that is favored towards Trial 99 where the users who rated an improvement were more likely to give a high overall rating. The proportion of users who found their sleep quality to be excellent is higher in Trial 99, and therefore can assume that the amount of the ingredient that is affecting sleep quality is higher in concentration in Trial 99 and should be kept at this concentration for future versions. In conclusion, Trial 51 outperforms in these various metrics by the proportion of users who found an improvement and is the better version of the supplement.