ANALYZING THE INTERACTION BETWEEN RTLX AND SUS SCORES IN VOICE USER INTERFACES: A STUDY FOCUSED ON AMAZON ALEXA

ANJALI RAJARAM NAYAK

Msc INFORMATION SYSTEMS UNIVERSITY COLLEGE DUBLIN

Abstract

This study looks at how workload and usability are related when people use voice user interfaces like Amazon Alexa via an Echo Smart Speaker. We had 100 participants perform 10 common tasks using Alexa on an Echo Smart speaker. Afterward they filled out two questionnaires: the Raw NASA Task Load Index to measure how much effort and stress they felt while using Alexa and the System Usability Scale to rate how easy they found it.The main goal of the study is to find out if there is a significant link between how hard people think a task is and how easy they find to use the system. We used statistical analysis to explore this connection. The results will help us understand how workload affects usability and guide the design of better voice interface

Results

The study explores the hypothesis

H0: There is no statistically significant relationship between RTLX and SUS

H1: There is a statistically significant relationship between RTLX and SUS.

Statistical data analysis was carried out to determine any relationship between Workload (RTLX) and Usability (SUS).Initial analysis of the 100 data points to check for any invalid data(null values,strings etc), showed that one value was below the minimum required score i.e 0 and one value was exceeding the value of 100 for SUS score. The scale is supposed to range from minimum of 0 to maximum of 100. Data cleaning was performed to get rid of the scores that didn't fit into this range so as that it won't be a problem to get the actual accurate result . This resulted in a dataset with 98 observations.

In the RTLX scores the scale is supposed to range from minimum 0 to maximum 126.All the scores in the dataset are within this range.

Hence **Descriptive Statistics** for the cleaned dataset was found , i.e mean,median,Standard deviation, IQR and minimum and maximum of scores.

SL No.	Statistics	SUS Score	RTLX Score
1	Minimum	0	20
2	Maximum	100	62
3	Mean	53.7	42.59
4	Median	52.50	42.50
5	Standard Deviation	23.32	9.37
6	Interquartile Range	34.37	12.75

TABLE 1: Descriptive Statistics for SUS and RTLX scores

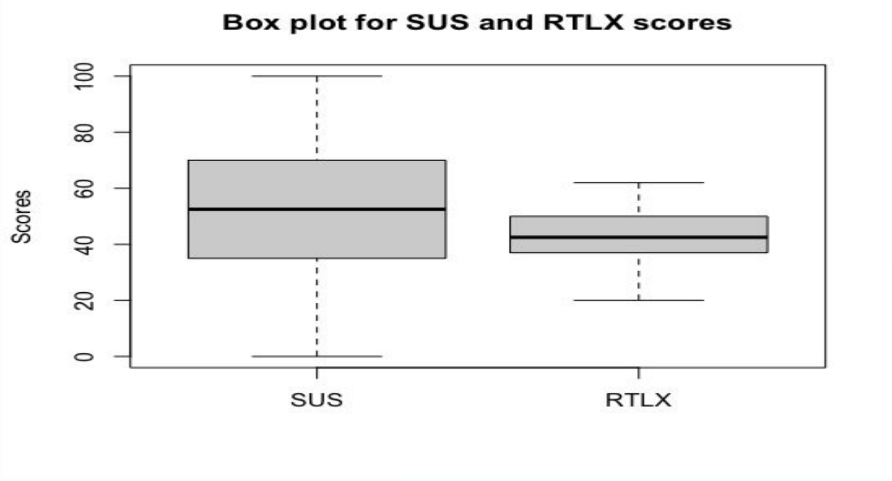


Figure 1. Boxplot for RTLX and SUS Scores.

The Box plot in figure one shows the average score and range of both SUS and RTLS,giving a clear view of how the scores are spread out.Also shows that there are no outliers.

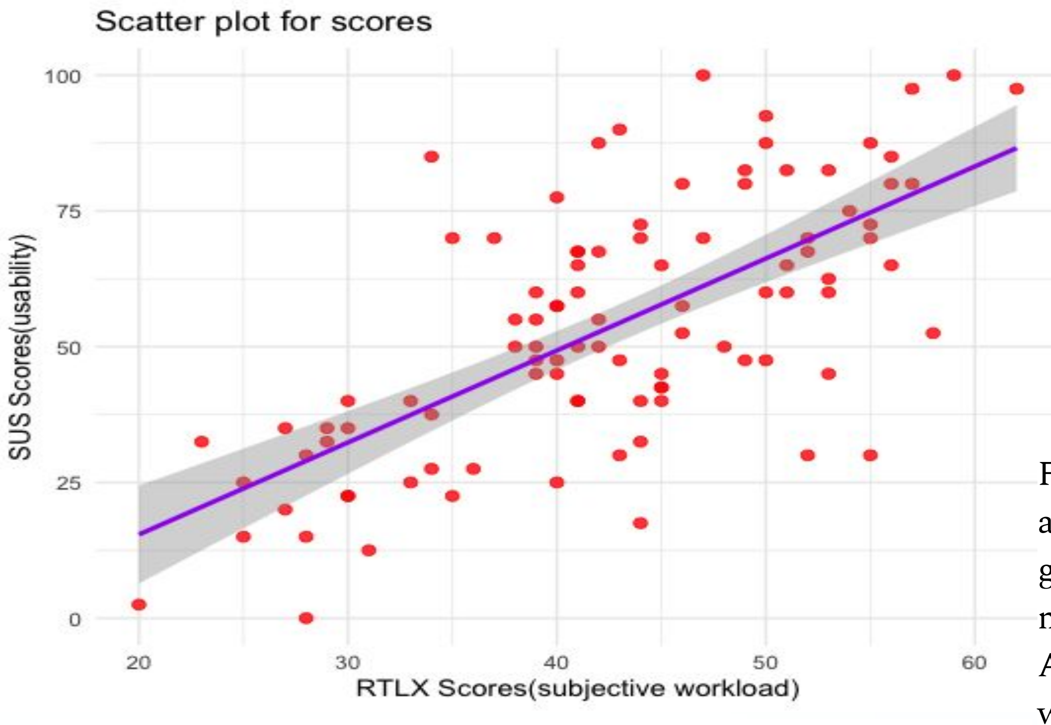


Figure 2: Scatter plot for SUS Vs RTLX scores

Scatter plot was generated to better understand the relationship between SUS and RTLX scores(Figure 2). A **linear relationship**(positive correlation) seemed to exist between SUS and RTLX scores.A line of regression is fitted which gives the positive relationship between RTLX and SUS scores

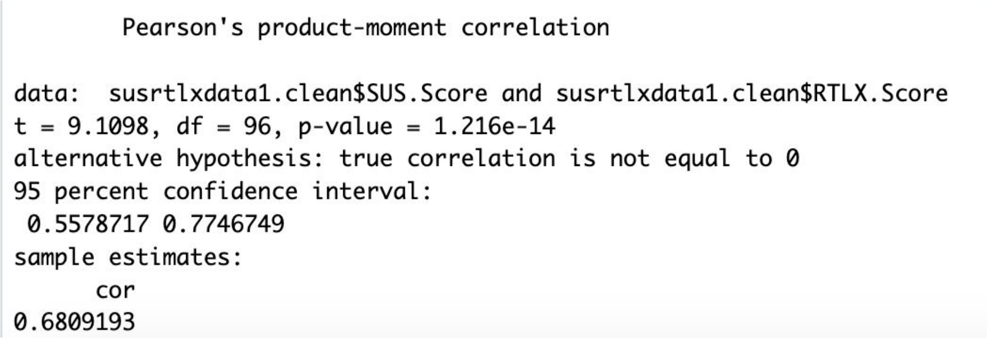
Correlation Analysis

After satisfying the assumptions,**Pearson's Correlation** test was conducted to statistically check the magnitude of the correlation between RTLX and SUS scores,i.e,correlation between Subjective workload and Usability.

The Pearson's correlation coefficient r can range from **+1** to **-1**.The closer it's value to either +1 or -1, the stronger is the association between the variables involved. The signs represents is the relation is positive or negative(Leards Statistics(2020)).

The test gave us the result **r(96)=0.68,p<0.001**.In this scenario moderately positive correlation is indicated by the r value **0.68** .The null hypothesis would be accepted if p value was greater than **0.001** and is rejected if p value is less than 0.001.(Beers 2020). As our p value is less than 0.001 we can conclude that **null hypothesis** is rejected and **H1** is accepted.

This depicts how as workload increases , usability also increases.



```
        Pearson's product-moment correlation

data:  susrtlxdata1.clean$SUS.Score and susrtlxdata1.clean$RTLX.Score
t = 9.1098, df = 96, p-value = 1.216e-14
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5578717 0.7746749
sample estimates:
      cor
0.6809193
```

Figure 3 :Correlation Test output

Here in the above figure **t**=test statistics, **df** =degree of freedom ((n-2) where n=number of rows in the dataset) and **p -value** gives the probability of the relationship occurring by chance. Hence in our case p value is **1.216e-14** which means the probability of relationship between RTLX and SUS is less than **0.001** .

Discussions

From Brooke J(1996)'s research,Usability says how spontaneous, efficient and how effective a system is for users. And System Usability Scale will give the quick and easy way when testing any interface which requires to measure it's usability.

And workload defines how a user should interact with a system .This workload is usually measured in NASA-TLX tools it's main feature is understand human behaviour in various aspects.(Hart, S. G., & Staveland, L. E. (1988))

Longo,L.,& Dondio P(2015) state that if the system requires more mental workload to use it then user will experience fatigue,frustration and likely to make some mistakes,which will lead to negative experience.Their study stated that even though workload and usability are directly proportional to user experience these aspects have their own features which may not always align with each other. But if put together the distinct features of workload and usability in future works we might be able to develop a more user friendly human interfaces. **This is contradicting our analysis as in our analysis RTLX and SUS are proved to be correlated.**

Wu, Y., Edwards, J., Cooney, O., Bleakley, A., Doyle, P. R., Clark, L., ... & Cowan, B. R. (2020), discuss on how non-native speakers tend to put more mental workload when using Interactive Personal assistance such as Amazon alexa and if this happens due to language barrier they may not be able convey their requirements efficiently which will directly affect the performance of the IPA

Limitations & Future Work

- There is no specific target group like gender,age,less educated or frequent users or people from various cultures to incur the result more accurately.
- There might be other personal factors that influence workload and usability measurement , which would not necessarily give accurate results.
- Usability and workload may depend on how complex the task is or user friendly, and if the study did take these into consideration results might vary(Longo,L.,& Dondio P(2015))

❖ Exploring how would usability and workload impact other under various constraints like time,complexity etc which would help in building user friendly and balanced interfaces.

❖ The study should broaden by considering people from different culture and how they will affect workload and usability which will help in creating interfaces used globally.

Conclusion

After performing the correlation test for subjective workload and usability from the p-value it was inferred that the H0(Null) hypothesis is rejected and H1 hypothesis true that RTLX score is statistically significant to SUS scores and there is moderately strong relationship between them .

We can conclude that as the usability of the system increases, subjective workload also increases.

References

[1] Laerd Statistics(2020).Pearson's product moment correlation.Statistical tutorials and software guides .Retrieved from https://statistics.laerd.com/statistical-guides/pearson-correlation-coefficient-statistical-guide.php

[2] Beers,B (2020,September 13).P-Value Definition.Investopedia.Retrieved from https://www.investopedia.com/terms/p/p-value.asp)

[3] Brooke, J. (1996). SUS-A quick and dirty usability scale. Usability evaluation in industry, 189(194), 4-7.

[4]Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index): Results of empirical and theoretical research. In Advances in psychology (Vol. 52, pp. 139-183). North-Holland.

[5] Longo, L., & Dondio, P. (2015). On the relationship between perception of usability and subjective mental workload of web interfaces. In 2015 IEEE/WIC/ACM International Conference on Web Intelligence and Intelligent Agent Technology (WI-IAT) (Vol. 1, pp. 345- 352). IEEE.

[6] Wu, Y., Edwards, J., Cooney, O., Bleakley, A., Doyle, P. R., Clark, L., ... & Cowan, B. R. (2020). Mental workload and language production in non-native speaker IPA interaction. In Proceedings of the 2nd Conference on Conversational User Interfaces (pp. 1-8).