Analysis Report for Significant Relationship between RTLX and SUS

Donghui Wang

University College Dublin

Author Note

Donghui Wang, School of Information & Communication Studies, University College Dublin.

Correspondence concerning this article should be addressed to Donghui Wang, School of Information & Communication Studies, University College Dublin, Dublin, D04 F6X4. E-mail: donghui.wang@ucdconnect.ie

Analysis Report for Significant Relationship between RTLX and SUS

Hypothesis and Background Knowledge

H_1: There will be a statistically significant relationship between RTLX and SUS.

NASA-TLX & RTLX:The NASA-TLX is a 6-item Likert scale questionnaire, measuring 6 constituent factors of mental workload: Mental Demand, Physical Demand, Temporal Demand, Performance, Effort, and Frustration. Scores on the questionnaire were summed to create an overall workload (Raw TLX) score (Wu et al.,2020).

SUS:The System Usability Scale (Brooke,1996) is a questionnaire that consists of ten questions. It is a highly cited usability assessment method and it has been massively applied (Bangor, Kortum and Miller,2008).

Result

This report was to explore statistical relationship between SUS and RTLX. Statistics data were collected from 100 participants took part in a study on the relationship between subjective workload and system usability in the use of voice user interfaces.

This section includes 3 parts: **I.Preparation, II.Data Cleaning and III.Statistical Analysis.**

I.Preparation

The packages and data were loaded. The data were save into SusRtlx data frame.

```r
library("dplyr")
library("ggplot2")
SusRtlx = read.csv("/Users/alex/Desktop/Homework/QDA/assignment1/SusRtlx.csv")
```

II.Data Cleaning

First, summary() function was used to inspect the data roughly.

```
summary(SusRtlx)
```

For SUS.Score, $maximum = 105$, $minimum = -2.50$. SUS.Score is from 0 to 100. RTLX.Score is from 0 to 120. The maximum of SUS.Score is larger than 100 and the minimum is less than 0. So there were outliers in the data frame.

Second, the function "which()" was used to located outliers.

```
which(SusRtlx$SUS.Score < 0 | SusRtlx$SUS.Score > 100)
```

```
## [1] 88 92
```

```
which(SusRtlx$RTLX.Score < 0 | SusRtlx$RTLX.Score > 120)
```

```
## integer(0)
```

2 impossible values were spotted. The ID indexes are 88 and 92. They were removed from SusRtlx in the following step via using filter().

```
SusRtlx = filter(SusRtlx, SusRtlx$SUS.Score >= 0 &
                    SusRtlx$SUS.Score <= 100)
```

Then RTLX.Score and SUS.score were saved into RTLX and SUS arrays. In this step, the medians,standard derivations, variances, means, upper bounds, lower bounds and IQRs were calculated and saved into variables.

```
RTLX = SusRtlx$RTLX.Score
SUS = SusRtlx$SUS.Score


RTLX_mean = mean(RTLX)
```

```
RTLX_var = var(RTLX)

RTLX_sd = sd(RTLX)

RTLX_upper_bound = RTLX_mean + 3*RTLX_sd

RTLX_lower_bound = RTLX_mean - 3*RTLX_sd

RTLX_mid = median(RTLX)

RTLX_IQR = IQR(RTLX)


SUS_mean = mean(SUS)

SUS_var = var(SUS)

SUS_sd = sd(SUS)

SUS_upper_bound = SUS_mean + 3*SUS_sd

SUS_lower_bound = SUS_mean - 3*SUS_sd

SUS_mid = median(SUS)

SUS_IQR = IQR(SUS)
```

3 standard derivation method was used to check the data again to make sure no outliers and impossible values in the data frame.

```
filter(SusRtlx,SusRtlx$RTLX.Score < RTLX_upper_bound &
       SusRtlx$RTLX.Score > RTLX_lower_bound &
       SusRtlx$SUS.Score < SUS_upper_bound &
       SusRtlx$SUS.Score > SUS_lower_bound)
```

The row number remained unchanged, so all the data in the data frame were normal.

III.Statistical Analysis

summary() function was used to assessed the data frame SusRtlx.

```
summary(SusRtlx)
```

Descriptive Statistics:

For SUS.Score: $minimum = 0$, $maximum = 100$, $median = 52.50$, $mean = 53.75$, $IQR = 34.375$, $stand\ derviation \approx 23.323$, $variance \approx 543.973$, $upper\ bound \approx 123.669$, $lower\ bound \approx -16.271$.

For RTLX.Score: $minimum = 20$, $maximum = 62$, $median = 42.50$, $mean = 42.59$, $IQR = 12.75$, $stand\ derviation \approx 9.371$, $variance \approx 87.811$, $upper\ bound \approx 70.704$, $lower\ bound \approx 14.480$.

Inferential Statistical Analysis:

cor() function was used to get Pearson's correlation coefficient and cor.text() function was used for correlation analysis.

```
r = cor(RTLX,SUS,method = "pearson")
cor.test(RTLX,SUS,method = "pearson")
```

Pearson's correlation coefficient is $r \approx 0.681$. The t value is $t \approx 9.110$. The degree of freedom is 96 and the p-value is $p = 1.216 * 10^{-14}$ and $p < 0.001$.

A statistically significant moderate positive correlation was discovered$[r(96) \approx 0.681, p < 0.001]$ between RTLX and SUS for the value of $p < 0.001$ and $t \approx 9.110$. So a positive relationship might exist between RTLX and SUS. The H_1 is supported by the data and the null hypothesis(H_0) is rejected.

Graph of the data frame SusRtlx:

ggplot() and geom_plot() were used to generate the graph. The positive relationship might exist as the rough schematic diagram showed.

```
ggplot(data = SusRtlx, aes(RTLX.Score,SUS.Score)) +

  geom_point(color="black") +

  geom_smooth(method = "lm",alpha = 0.2, color = "#1a9641") +

  labs(x = "RTLX Score", y ="SUS Score") +

  theme_apa(box = TRUE) +

  theme(legend.position = c(2, 2))
```
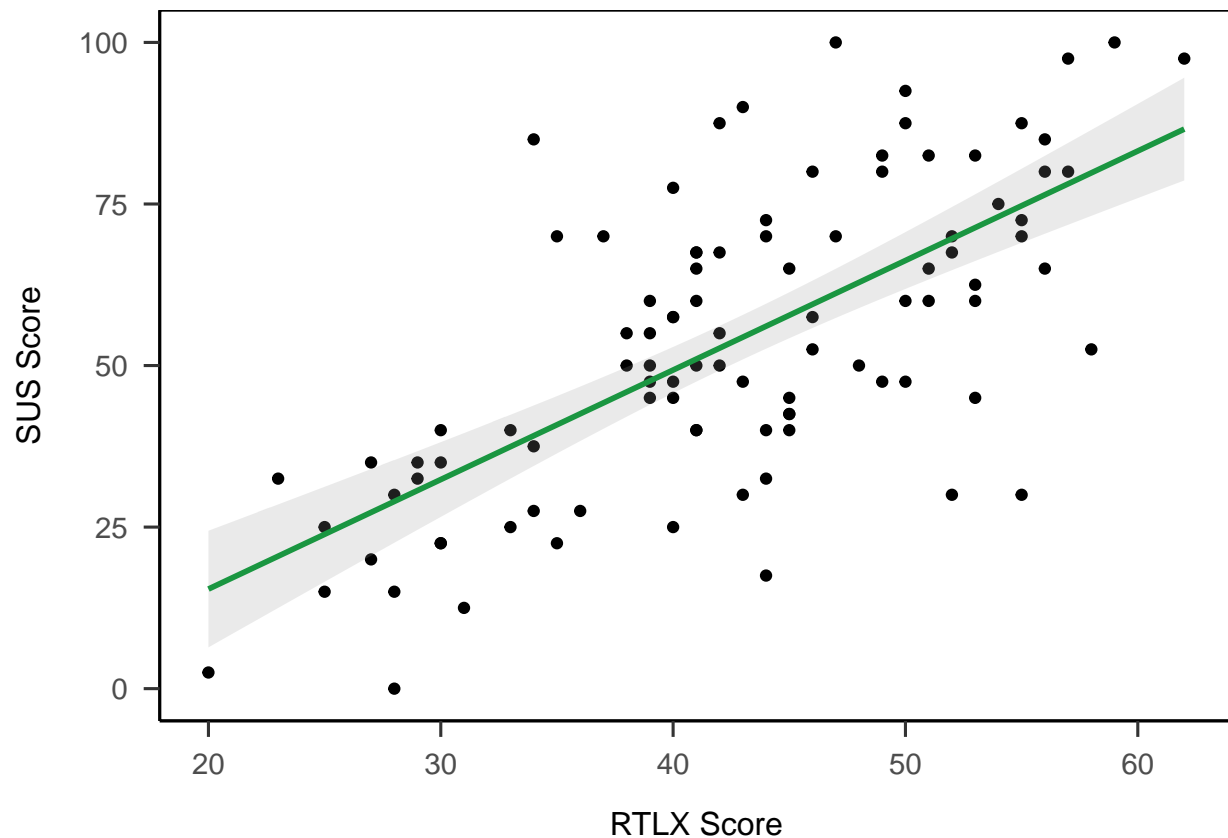

Figure 1. Distribution of RTLX and SUS

Discussion

I.Summary

This report explored the relationship between RTLX and SUS. After removing 2 outliers in the data frame, the left were reloaded.The inferential statistically analysis

revealed a statistically significant moderate positive relationship via Pearson Correlation. This result conflicted with Longo and Dondio's experiment (2015) and Longo's study (2017). As there was no repeated experiments and comparing group, the significant relationship need to be checked in further research.

II.Limitations

1.Detail information should be provided. Comparing to Longo and Dondio's experiment (2015) and the mixed-deigned experiment (Wu et al.,2020), there are no information like age, tasks and skill level. This research needs more detailed experiments to clarify whether factors are related to the relationship and how these factors influence it.

2.No repeated experiments and comparing groups. It's not enough to conduct a concrete conclusion about the significant relationship. The result is conflict with Longo and Dondio's study (2015) and Longo's study (2017) for the coefficient could be negative in their experiments.

III.Future Research

1.More groups could be designed as comparing groups to explore how a certain factor works. The comparing experiments could help to find how the factors will influence the result and diminish more uncertainty.

2.More experiments could be made in the same conditions in different time to make sure the result is not a coincidence and reduce the randomness.

IV.Conclusion

Though the data showed the statistically significant moderate positive relationship between RTLX and SUS, the result conflicted with Longo and Dondio's study (2015) and Longo's study (2017) in this area. Further research is needed.

Reference

Bangor, A., Kortum, P. T., & Miller, J. T. (2008). An empirical evaluation of the system usability scale. *Intl. Journal of Human–Computer Interaction, 24*(6), 574-594.

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

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.

Longo, L. (2017). Subjective usability, mental workload assessments and their impact on objective human performance. In *IFIP Conference on Human-Computer Interaction* (pp. 202-223). Springer, Cham.

Wu, Y., Edwards, J., Cooney, O., Bleakley, A., Doyle, P., Clark, L., Rough, D., & Cowan, B. (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).