An analysis of the relationship between the Raw NASA Task Load Index and the System Usability Scale

Yang Yinfang

University College Dublin

Results

Data Cleaning

Data cleaning was performed in order to correct or remove impossible scores and outliers from the data set to ensure reliable results. The scale of RTLX ranges from a minimum total score of 0 to a maximum of 126. The scale of SUS ranges from a minimum total score of 0 to a maximum total score of 100. In the SUS data, two impossible scores beyond the extremes were detected. This problematic data should be an incorrect impossible score rather than an outlier. It is not possible to know the the actual score or use this incorrect score to infer the true score. To avoid inferential statistical bias due to problematic data, these two figures in row 92 and row 88 were removed. The dataset has 100 rows of data and can afford the loss of two rows.

The variables were then visualised by histograms and boxplots. The boxplots indicated that there was no outlier on both variables. A better definition for outliers is that “three standard deviations away from the mean”. Therefor the data and the mean plus/minus three standard deviations were compared. The comparison resulted in no outliers.

The the normal distribution were visually assessed. Data are normally distributed without skew.

Descriptive statistics

The sample size is 98 rows of data after the date cleaning.

Table 1: Descriptive statistics for RTLX and SUS score

	Mean	Standard deviation	Median	IQR
SUS.Score	53.70	23.32	52.5	34.38
RTLX.Score	42.59	9.37	42.5	12.75

The actual data for SUS.Score has a minimum value of 0 and a maximum value of 100. The actual data for RTLX.Score has a minimum value of 20 and a maximum value of 62.

Inferential Statistics

Pearson correlation coefficient was calculated, which is the most common way of measuring a linear correlation between two variables and can be used to test statistical hypotheses. The experimental hypothesis H1 is the following: There will be a statistically significant relationship between RTLX and SUS.

The Pearson correlation coefficient value provided evidence of a strong positive correlation between RTLX and SUS. $r(96) = 0.68$, $p < 0.001$ (Figure 1). The findings were statistically significant. And H0 is rejected and H1 is accepted.

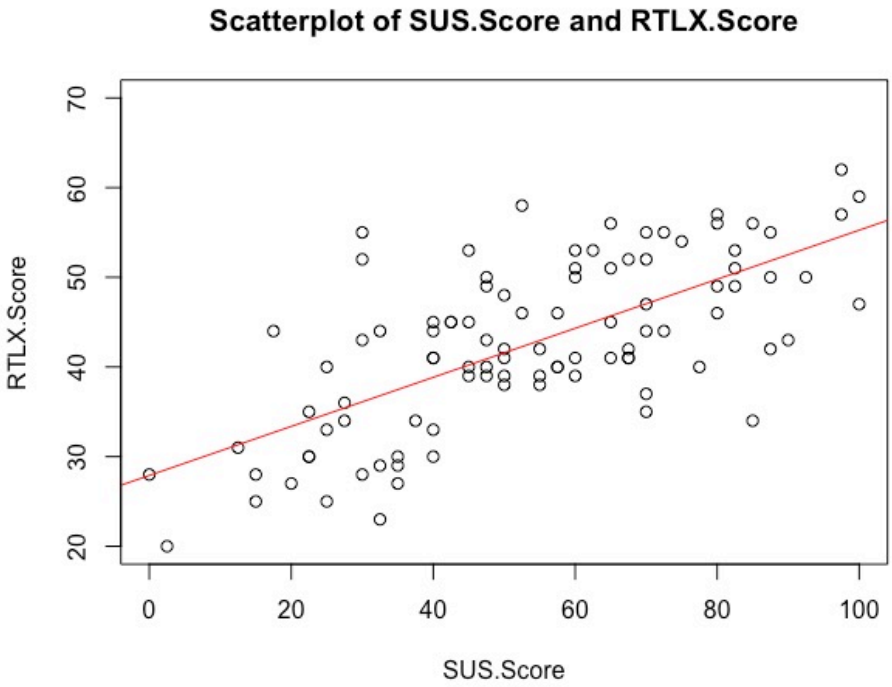

Figure 1: Scatterplot of SUS.Score and RTLX.Score

Discussion

The study aimed to identify the relationship between subjective workload and system usability in the use of voice user interfaces. Subjective workload were measured by the Raw NASA Task Load Index. System usability were measured by the System Usability Scale. A statistically significant strong positive correlation between RTLX and SUS was found. Positive correlation refers to the relationship formed

between the subjective workload and the system usability where both of them move in a similar direction.

Researchers in related fields are as well studying the correlation between the subjective workload and the system usability, or have conducted systematically designed studies using the two questionnaires. However, there are fewer relevant studies and this positive correlation has not been widely validated. The human mental workload (MWL) mentioned in Longo and Dondio's research is a similar concept to the subjective workload of this study. They refer to it as a foundational concept for the study of human interaction with computers and other technological devices. It has been widely documented that mental overload or underload can have a negative impact on performance(Xie & Salvendy, 2000). A low level of MWL can lead to annoyance and frustration when people are processing information. A high level of MWL situations can be problematic or even dangerous, leading to confusion, reducing efficiency in information processing and increasing the chance of errors and mistakes. (Longo & Dondio, 2015)

The RTLX index, which measured subjective workload for users in this study, has an extreme range of 0 to 126, but the actual dataset has an extreme range of 20 to 62, which may indicate that the participants in this experiment were in a moderate state of subjective workload and did not experience overload or underload. The actual and theoretical ranges of SUS.Score are both 0 to 100. Thus for this Amazon Alexa via an Echo smart speaker system, subjective workload and system usability in the use of voice user interfaces present a strong positive correlation.

Longo and Dondio's study found no clear relationship between the perception of usability of a set of web-interfaces and the mental workload. " Specifically, a well known subjective instrument for assessing usability —the System Usability Scale — and two subjective mental workload assessment procedures —the NASA Task Load Index, and the Workload Profile —have been employed in a user study involving 40 subjects. Empirical evidence suggests that there is no clear re- lationship between the perception of usability of a set of web-interfaces and the mental workload imposed by a set of designed tasks to be executed on them." However, part of the task from this study showed a moderate negative correlation, a high positive correlation and a high negative correlation. A moderate negative correlation may be explained by the fact that the perception of good usability of the task interface increases when the attention required for the task is moderately reduced. A high positive correlation may be explained by the fact that perceptions of good usability are enhanced if the task is enjoyable, even if the amount of mental effort increases. A high negative correlation may be explained by the fact that the perception of usability is severely negatively affected by an increase in the user's mental workload and by the fact that the task is not easy(Longo & Dondio, 2015).

Hart, in a 20-year follow-up study of the NASA- TLX questionnaire, stated that the difficulty of the task as a key factor in relevance. "NASA- TLX ratings may or may not covary with measures of performance (dissociation). For example, the cost of performing well in a difficult task may be an unacceptably high level of workload. On the other hand, the workload cost of performing an apparently undemanding

vigilance task can be extremely high prompted by boredom unless ameliorated by improved design. ”(Hart, 2006).

Limitations

An obvious limitation is that the experiment was conducted as a cross-sectional study only, with a representative sub-group of data analysed at a given point of time. The advantage of the experiment is that it avoids the variation associated with extracting data from different time points, from different populations. It does not involve manipulation of variables. This type of study is a descriptive and observational study that cannot accurately obtain causal relationships and can only be used to infer possible relationships. This study provides a useful stepping stone for further research. In future studies, longitudinal studies can be conducted, where the study population is studied and observed more than once. Researchers can observe how variables change over time. However, it is important to take into account the speed of industry updates and evolutions to ensure consistency of the product or product version.

The experiment requires more careful selection of participants to control for more variables. The study is about the system usability in the use of voice user interfaces, which should take into account the language ability of the participants, the type of native language of the participants and whether the participants participate in the experiment in their native language or not. A study shows that whether or not users conducted the experiment in their native language significantly affected mental workload. “We present a mixed-design experiment, where native (L1) and non-native

(L2) English speakers completed tasks with IPAs via smartphones and smart speakers. We found significantly higher mental workload for L2 speakers in IPA interactions." (Wu et al., 2020). The experience of the participants can significantly influence the evaluation of system usability."Users having a more extensive experience with a product tended to provide higher, more favorable, SUS scores over users with either no or limited experience with a product—and by as much as 15-16%, regardless of the domain product type."（McLellan et al., 2012). These variables need to be controlled or recorded in the next experiments.

This experiment is based on self-reported data, and self-reported data may contain several potential sources of bias. The advantage is that the participants give a more realistic response in anonymity. But participants may not be able to assess themselves or explain the questions accurately. However, self-reporting is common in relevant studies. "Usability does not exist in any absolute sense; it can only be defined with reference to particular contexts. This, in turn, means that there are no absolute measures of usability."(Brooke, 1996).

Finally, attention to ecological validity is needed, and experiments can be conducted both inside and outside the laboratory. Laboratory-based studies can reduce potential impacts such as noise and distraction. It also allows users to be aware that they are being recorded (Wu et al., 2020). A large amount of RTLX experiments are also carried out outside the laboratory. "The majority of the studies targeted a specific operational environment, even though the actual study was performed in a laboratory (10%) or simulation environment."(Hart, 2006).

Conclusion

Using data collected by two questionnaires, the Raw NASA Task Load Index (RTLX - Hart & Staveland, 1988; Hart 2006) and the System Usability Scale (Brooke 1996), a statistically significant strong positive correlation can be found between the subjective workload and system usability in the use of voice user interfaces.

References

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

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.

Hart, S. G. (2006). NASA-task load index (NASA-TLX); 20 years later. In Proceedings of the human factors and ergonomics society annual meeting (Vol. 50, No. 9, pp. 904-908). Sage CA: Los Angeles, CA: Sage publications.

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.

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).

Xie, B., & Salvendy, G. (2000). Review and reappraisal of modelling and predicting mental workload in single- and multi-task environments. *Work & Stress*, *14*(1), 74–99. https://doi.org/10.1080/026783700417249

McLellan, S., Muddimer, A., & Peres, S. C. (2012). The effect of experience on system usability scale ratings. Journal of usability studies, 7(2), 56-67.