LOAD AND CODE: BALANCING SYSTEM USABILITY AND SUBJECTIVE WORKLOAD IN VOICE MODE

Insights from the Raw NASA Task Load Index and System Usability Scale in Amazon Alexa

AUTHORS

SAI SUGUN D R
MSc INFORMATION SYSTEMS
UNIVERSITY COLLEGE DUBLIN

AFFILIATIONS

UNIVERISTY COLLEGE DUBLIN

ABSTRACT

This study looks at the relationship between subjective workload and system usability in voice user interfaces, namely Amazon Alexa. A total of 100 participants did ten daily activities with Alexa using an Echo smart speaker, following which they completed the Raw NASA Task Load Index (RTLX) (Hart & Staveland, 1988; Hart, 2006) and the System Usability Scale (SUS) (Brooke, 1996). The RTLX assesses subjective workload using six dimensions: mental demand, physical demand, temporal demand, performance, effort, and frustration, whereas the SUS assesses subjective system usability. The findings demonstrated a strong positive relationship between increased workload and decreased usability, implying that as effort grows, system usability perception declines. This understanding of the relationship between cognitive demands and usability emphasizes the need for design improvements in voice user interfaces to minimize burden and increase user experience. These findings are consistent with previous studies on workload and human-computer interaction (Wu et al., 2020), highlighting the importance of knowing task complexity in defining system usability. The report also includes concrete ideas for lowering workload and improving the general usability of speech technologies.

OBJECTIVE

- To investigate the relationship between subjective workload, as evaluated by the Raw NASA Task Load Index (RTLX) (Hart & Staveland, 1988; Hart, 2006), and system usability, as measured by the System Usability Scale (SUS) (Brooke, 1996), in voice user interfaces, namely Amazon Alexa.
- To uncover usability issues that develop when cognitive effort increases, user's experiences with voice assistants during everyday tasks were examined.
- To make practical recommendations for improving voice user interfaces while reducing workload and improving overall user experience.

RELATED LITERATURE

- Brooke, J. (1996). SUS - A quick and dirty usability scale. [https://www.researchgate.net/publication/228593520]
- Hart, S. G., & Staveland, L. E. (1988). Development of NASA-TLX (Task Load Index). [https://doi.org/10.1016/S0166-4115(08)62386-9]
- Hart, S. G. (2006). NASA-Task Load Index (NASA-TLX); 20 Years Later. [DOI: 10.1177/154193120605000909]
- Wu, Y., et al. (2020). Mental Workload and Language Production in Non-Native Speaker IPA Interaction. [DOI: 10.1145/3405755.3406118]

METHODOLOGY

Participants: The study included 100 people of varied ages, genders, and technological expertise. To test interaction with voice user interfaces, each participant completed 10 predefined everyday tasks using Amazon Alexa on an Echo smart speaker.

Task Procedure: Participants followed a standardized list of daily actions, including setting reminders and operating smart home devices, to ensure a consistent engagement experience for all users. The tests were created to accurately assess usability and workload by simulating real-life circumstances and reflecting common Alexa applications.

Assessment Tools: After completing activities, participants filled out two surveys. The Raw NASA Task Load Index (RTLX) (Hart & Staveland, 1988) assessed six characteristics of subjective workload: mental, physical, and temporal demands, performance, effort, and frustration. The System Usability Scale (SUS) (Brooke, 1996) offered information about perceived system usability. Both scales were used immediately to obtain genuine user perceptions.

Data Analysis: Post task replies were gathered via internet platforms. Descriptive statistics were provided for both RTLX and SUS scores. Pearson's correlation coefficient was used to determine the relationship between workload and usability. Using the R programming language, data visualization techniques such as scatterplots, histograms, and boxplots were used to demonstrate score association and distribution patterns.

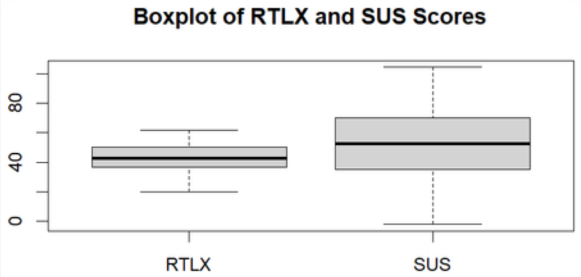

Figure 1. Boxplot for RTLX and SUS Scores.

RESULTS

SL.NO	STATISTICS	SUS SCORE	RLTX SCORE
1.	MEAN SCORE	42.62	53.65
2.	MEDIAN SCORE	42.5	52.5
3.	STANDARD DEVIATION	9.41	24.32
4.	MAXIMUN SCORE	62.0	105
5.	MINIMUN SCORE	20	-2.5
6.	Q1(25%)	36.75	35
7.	Q2(75%)	50	70

TABLE I. Descriptive Statistics for RTLX and SUS Scores.

The table shows descriptive statistics on the System Usability Scale (SUS) and Raw NASA Task Load Index (RTLX) scores of the participants. The average SUS score of 42.62 indicates a generally acceptable voice interface, while the median is 42.5, indicating consistency in usability opinions. The mean RTLX score of 53.65 and the median of 52.5 indicate a moderate level of subjective workload. The standard deviations for SUS and RTLX are 9.41 and 24.32, respectively, indicating that user experiences vary. The greatest SUS score is 62.0, while the lowest is 20, indicating a variety of usability perceptions. In terms of workload, RTLX ratings range from -2.5 to 105, with quartiles indicating that 25% of participants assessed usability below 36.75 and experienced workload below 35. 75% of respondents rated usability below 50 and subjective workload below 70. Overall, these findings show that, while the voice interface is widely regarded as usable, participant's workload experiences vary greatly.

CONCLUSION

I conclude that, in my opinion, the study demonstrates a significant relationship between subjective workload and system usability in voice user interfaces like Amazon Alexa. The positive correlation between RTLX and SUS scores, as captured by Pearson's correlation coefficient **(r = 0.69, p < 0.001)**, indicates that as user's perceived workloads increase whether mental, physical, or temporal their perception of the system's usability decreases. This supports the importance of minimizing cognitive demands to enhance user experience. The findings from the mean **SUS score of 42.62** and **mean RTLX score of 53.65** further illustrate that moderate levels of workload correlate with lower usability ratings. These insights provide a clear direction for designing more intuitive and user-friendly voice interfaces. Ultimately, reducing workload can lead to smoother and more engaging interactions with voice technologies, improving overall user satisfaction.

ANALYSIS

The examination of RTLX and SUS ratings offers vital information about the relationship between subjective workload and system usability in voice user interfaces. As shown in the boxplot, RTLX ratings are less variable, implying that participant's subjective workload was rather consistent throughout jobs conducted using Amazon Alexa. In contrast, the broader range of SUS scores represents a more diversified set of perspectives on system usability. The scatterplot demonstrating a positive association between RTLX and SUS scores indicates that as workload increased, usability ratings decreased, which is consistent with earlier research on the influence of mental and physical effort on system interaction (Hart & Staveland, 1988; Brooke, 1996). This link is consistent with research on mental effort and system performance, which shows that increased workload frequently reduces a system's subjective usability (Wu et al., 2020). These findings highlight the importance of balancing burden and usability when designing user interfaces for smart devices, especially those used for daily chores. Incorporating usability and workload indicators, such as NASA-TLX and SUS, is critical for understanding how users perceive and engage with voice-activated devices (Hart, 2006).

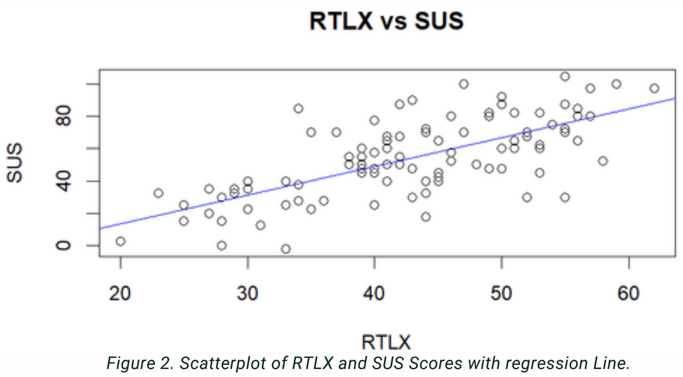

Figure 2. Scatterplot of RTLX and SUS Scores with regression Line.

Figure 2 shows a scatterplot of the positive linear association between RTLX and SUS scores, with the fitted regression line indicating a clear increasing trend. Higher subjective workload (higher RTLX scores) is linked to worse usability evaluations (lower SUS scores) (correlation coefficient = 0.689, p < 0.001). This finding is consistent with previous research indicating that increased task load frequently coincides with a decrease in subjective system usability.

Figure 3 provides vital insights into the variability in participant's subjective workload and subjective system usability during their interactions with Amazon Alexa. The histograms show a more concentrated distribution for RTLX scores, with a peak in the 40-50 range, indicating that the majority of participants had moderate workloads. In contrast, SUS scores have a greater distribution, with some participants ranking usability very low and others very high, resulting in a wider range of results. This variation may be due to variances in user familiarity with speech interfaces or personal preferences for interaction methods.

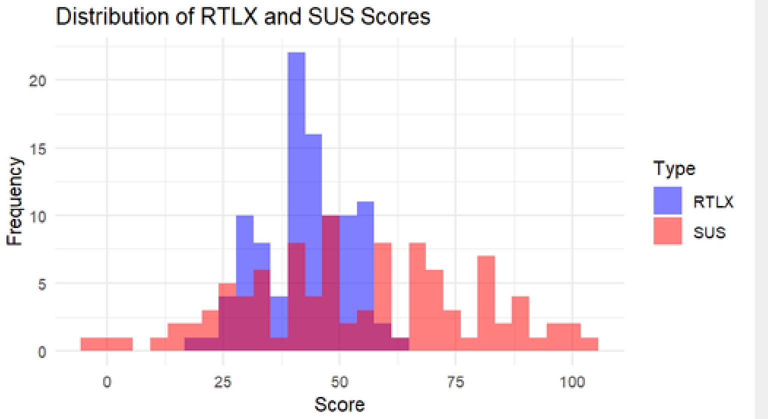

Figure 3. Distribution of RTLX and SUS Scores