CHAOS OR CLARITY

• Unraveling the SUS-RTLX • Riddle in Voice User Interfaces

D.S.D.Hemanth(23202425)
University College Dublin

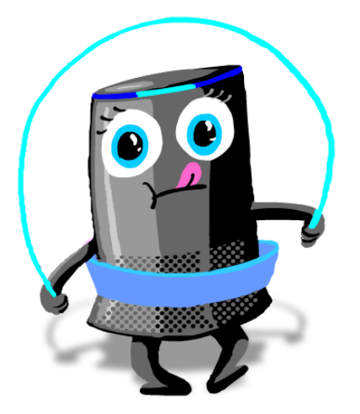
Introduction

***Background*:** Voice assistants are everywhere, but do they make life easier or add to the daily chaos?

***Significance*:** Our study aims to uncover the secrets behind users' feelings about workload and usability when it comes to these talkative gadgets.

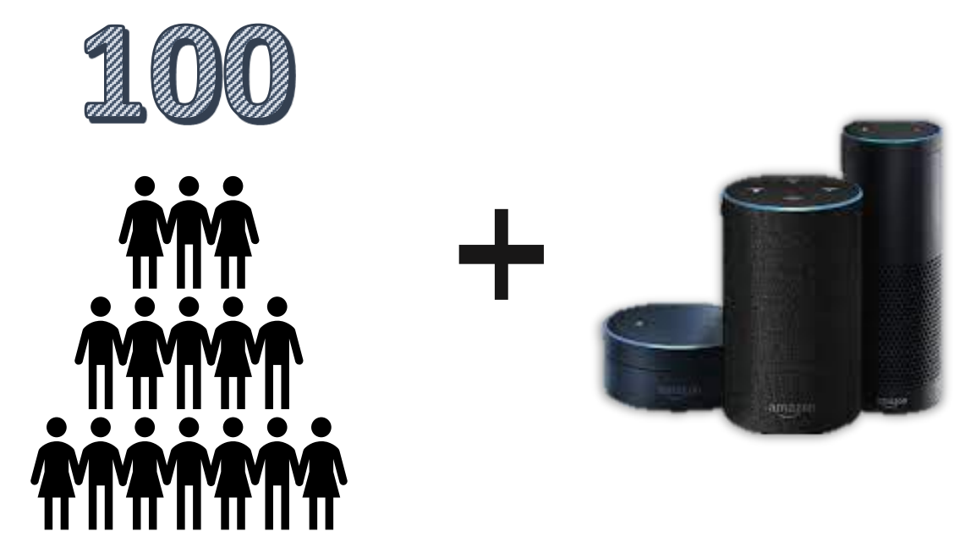
***Question*:** Is there a clear connection between how users rate the effort (RTLX) and the Usability (SUS) of their voice assistants, or is it all just tech talk? (H1)

Common Assumption: Voice assistants are designed to simplify our lives. Therefore, it's reasonable to expect that as their usability improves, the workload should decrease, aligning with the goal of creating a helpful personal assistant. (Consider as H2)



Methodology

Participants:



→ 10 everyday tasks using voice commands.

Materials:

Our trusty tools included the Raw NASA Task Load Index (RTLX) questionnaire, which assessed workload across dimensions, and the System Usability Scale(SUS) to measure the usability of the voice assistant.

RTLX examined six essential elements: mental demand(MD), physical demand(PD), temporal demand(TD), performance(Pe), effort(Ef), and frustration(Fr). The provided images offer a visual representation of the weighted subscale ratings and an overall workload(OW) value. (RTLX - Hart & Staveland, 1988; Hart ,2006)

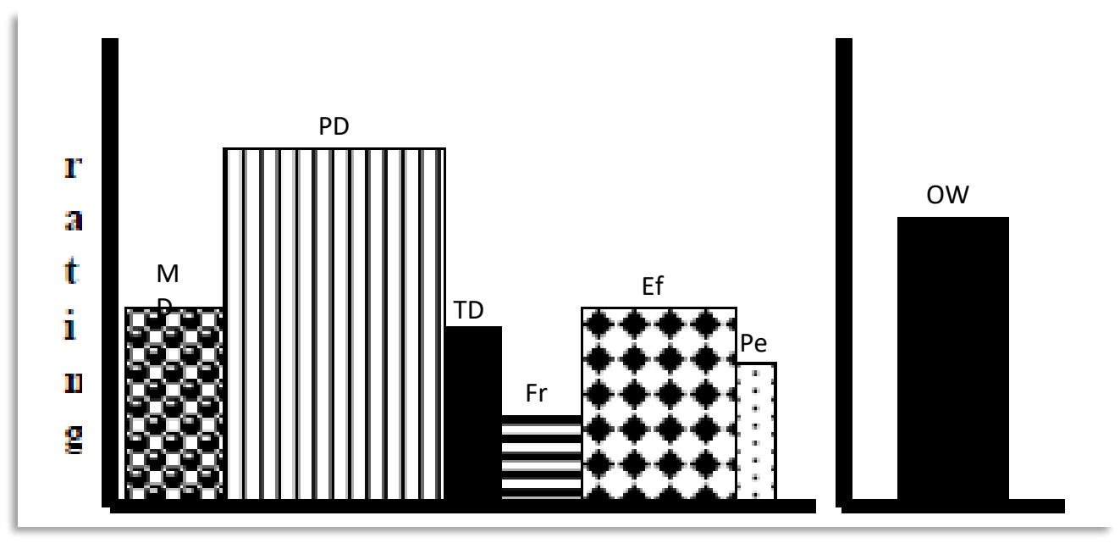


Figure 1: Graphic Representation of weighted subscale ratings and an overall workload value(Hart ,2006)

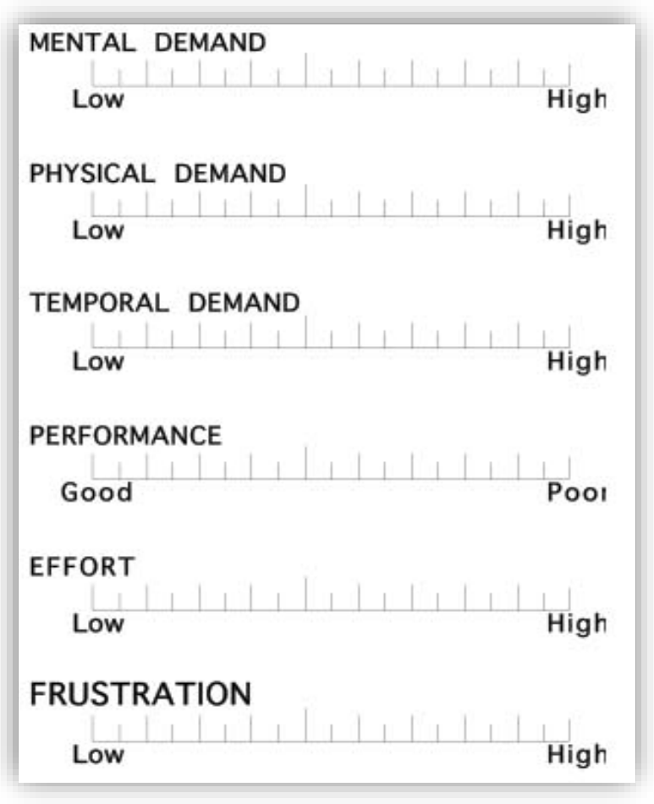

MENTAL DEMAND
Low
High
PHYSICAL DEMAND
Low
High
TEMPORAL DEMAND
Low
High
PERFORMANCE
Good
Poor
EFFORT
Low
High
FRUSTRATION
Low
High

Figure 2: Six subscales (Hart ,2006)

SUS consists of ten items or questions rated on a five-point Likert scale, ranging from 0 to 4. (Brooke 1996)

Procedures:

Once participants successfully completed their tasks, they thoughtfully responded to both the RTLX and SUS questionnaires, offering valuable insights into their perceptions of workload and usability. Subsequently, proceeded with a meticulous analysis of the questionnaire data to unveil the intricate dynamics of voice assistants and their implications for everyday routines.

Data Cleaning:

Data cleaning was essential to maintain data quality and integrity. We encountered anomalies in our System Usability Scale (SUS) data. These anomalies included negative SUS scores, which are not within the standard SUS scoring range of 0 to 100, as well as a SUS score that exceeded 100.

To address these issues, we employed a data transformation approach, bringing all SUS scores within the valid range (0-100). This ensures that our analysis is based on consistent and reliable data. This keeps the dataset on a common scale, and maintain the interpretability of extreme values.

Hey Alexa, take out the garbage for me?
Um, I don't have arms & legs, Steve.

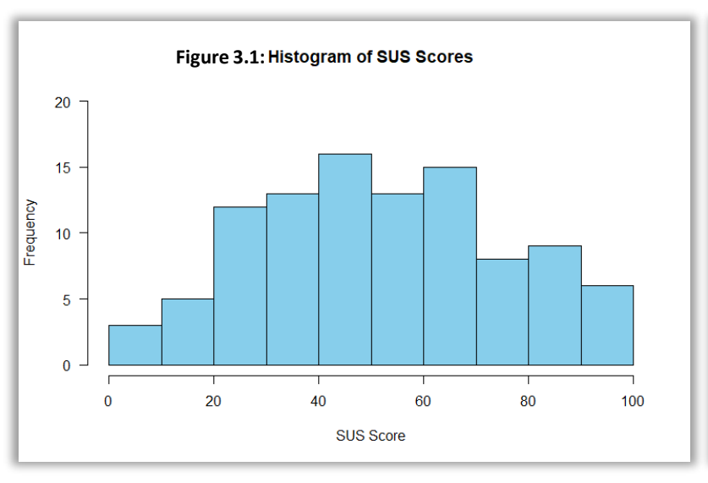
Results

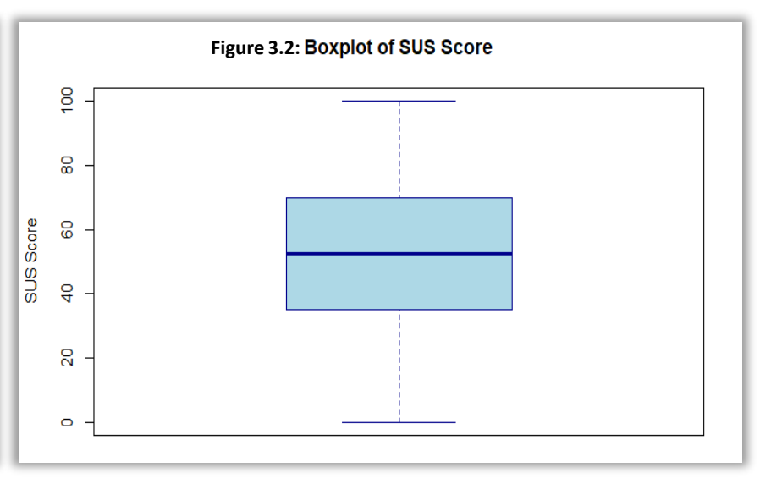
Table 1: Descriptive Statistics

Measure	SUS Scores	RTLX Scores
Mean	53.63	42.62
Standard Deviation	24.16	9.41
Maximum	100	62
Minimum	0	20
Median	52.50	42.50
Interquartile Range (IQR)	35	13.25

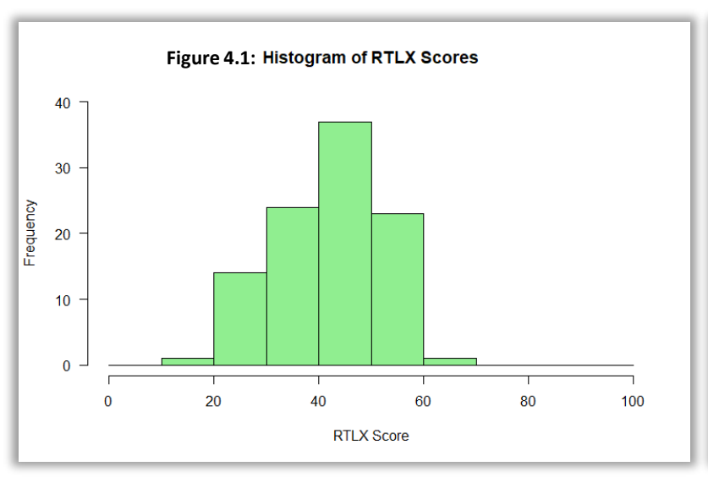
Data Distribution:

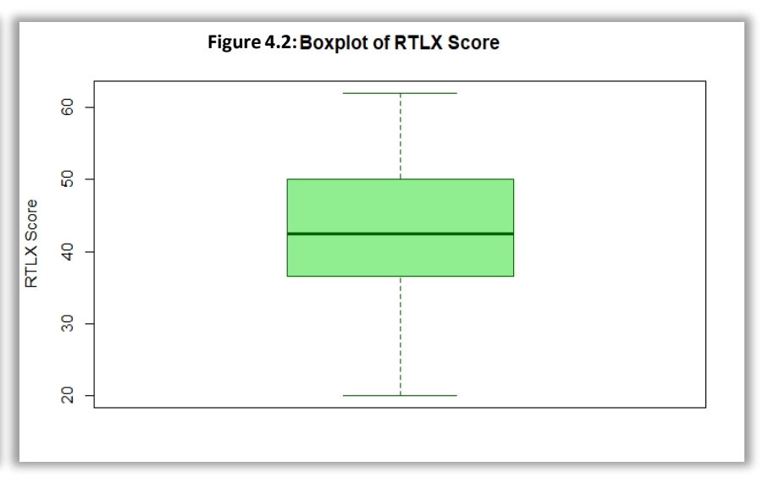
System Usability Scale:





RTLX:





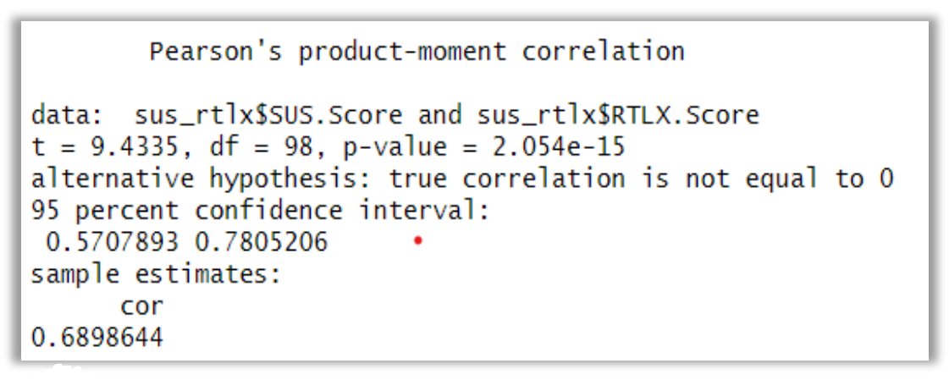
Observations:

- The SUS score histogram(Figure 3.1) reveals several peaks(a multimodal distribution), signifying diverse subgroups (e.g., variations among individuals and external factors) rather than a single distribution.(Jamie DeCoster, 2005)
- There are no potential outliers that can cause problems in our analysis.

Relationship Between SUS and RTLX Scores:

- We have performed Pearson's correlation test, and the correlation coefficient is approximately 0.6899, indicating a strong positive correlation. (H1)
- A significant positive correlation ($r(98) = 0.6899$, $p < 0.001$) between SUS and RTLX scores reveals that as voice assistant usability (SUS) increases, perceived workload (RTLX) also tends to rise. (might not be the case)
- Our research gave us an unexpected outcome that goes against the H2. We found that when voice user interfaces are more user-friendly, they tend to be linked with a higher sense of workload. This surprising finding requires further investigation to uncover the reasons behind it.

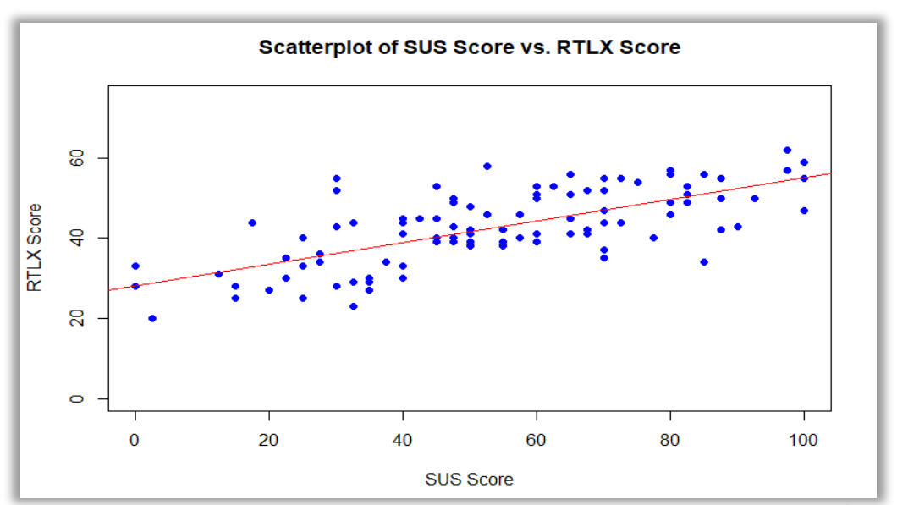
Correlation test Result



```
        Pearson's product-moment correlation

data:  sus_rtlx$SUS.Score and sus_rtlx$RTLX.Score
t = 9.4335, df = 98, p-value = 2.054e-15
alternative hypothesis: true correlation is not equal to 0
95 percent confidence interval:
 0.5707893 0.7805206
sample estimates:
      cor
0.6898644
```

Figure 5: scatterplot of SUS scores vs. RTLX scores



Observations:

- Non-random strong positive correlation.
- This suggests that an increase in SUS scores may, at times, coincide with an increase in workload, though this observation might not always hold true. Conversely, there is a possibility that higher workload levels could lead to improved usability.

Example:

A person may hesitate to use assistance for simple tasks they can easily handle on their own. However, assistance becomes valuable when they perceive a significant workload.

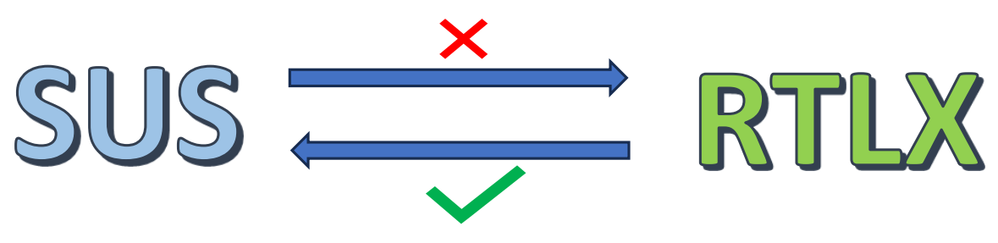
RTLX and SUS scores exhibit variability. Just as one might opt for a forklift for heavy loads, an increase in SUS scores doesn't always coincide with a rise in workload. Conversely, high workloads may lead to improved usability. However, it's important to consider the influence of errors and other factors, which we will discuss further.

Discussion

Our findings highlight a complex relationship between RTLX and SUS scores, challenging any simple cause-and-effect conclusions. The discussion that follows delves into these intricacies, seeking to unravel the underlying reasons and their broader implications.

People often mistake correlations as proof of causation. "Remember, while correlation is a necessary element for establishing causation, it's not the only factor." (Zaniletti et al., 2022). At least part of the observed association between two variables may arise by reverse causation or by the confounding effect of a third variable. (Alberto Abadie, 2005; Zaniletti et al., 2022)

- Random Chance: Not a Factor - High Significance ($p < 0.001$) Confirms a Meaningful Correlation
- ***Reverse causality***, a potential key factor, may explain our unexpected results, where increased workload (RTLX) affects higher SUS scores, suggesting a bidirectional influence. (Zaniletti et al., 2022)



- In research, third extraneous variables can serve as 'colliders,' 'mediators,' or 'confounders,' each influencing outcomes in distinct ways.(MacKinnon & Lamp, 2021)

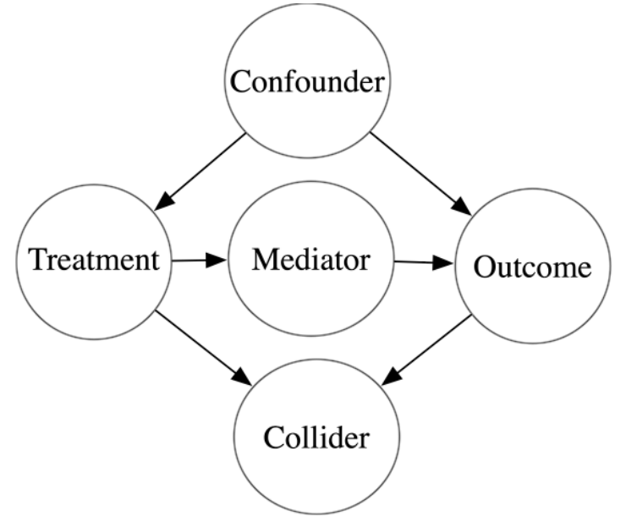


Figure 6: A diagram showing common causal relationships.(Keith et al., 2020)

- A Confounder is an extraneous variable whose presence affects the variables being studied so that the results do not reflect the actual relationship between the variables under study. (Pourhoseingholi, Baghestani, & Vahedi, 2012)."

Environment: Noise and lighting may impact scores.
Task Complexity: Varying task difficulty can influence results.
Order Effects: Task sequence could affect participant ratings.
Response Variability: Participants interpret and answer differently.
Time Changes: Study events may impact variables.
Additional Variables: Other factors (age, experience, cognitive abilities) can influence scores, strengthening the correlation.

Controlling confounding is essential for reliable research. Methods like randomization, stratification, and instrumental variables help ensure accurate, trustworthy results.(Wunsch, G. (2007)) (Pourhoseingholi et al., 2012).

- **Randomization:** Shuffling task order reduces unknown variable impact, revealing the true SUS-RTLX relationship.
- **Stratification:** Segment data by characteristics (e.g., age, experience) to unveil hidden patterns and identify influential subgroups.
- **Instrumental Variables**: Isolating causal links with specific variables strengthens the SUS-RTLX connection while controlling for confounding factors.

Is SUS an appropriate tool for evaluating the usability of voice interfaces?

"The widely used SUS scale might not be the most suitable choice for assessing voice assistants' usability since it was originally designed for graphical user interfaces (GUIs)."(Zwakman et al. (2021).

Further enhancements are necessary for the SUS scale to obtain more accurate usability measures or the creation of a new scale designed specifically for voice interfaces, such as the **VUS** (Voice Usability Scale). (Zwakman et al. (2021).

Overview:
Complex RTLX-SUS relationship: Correlation ≠ Causation. Factors: Reverse causality, confounders, extraneous variables, order effects, participant variability. Control with randomization, stratification, and instrumental variables. SUS suitability for voice interfaces questioned; VUS scale suggested.

Limitations of our study:
1.Sample Size
2.Demographic Variation
3.External Factors
4.Reverse Causality
5.Response Bias
6.No Mediator Analysis

Conclusion

Our study uncovered a fascinating correlation between SUS and RTLX scores. This finding implies a more intricate, possibly bidirectional connection, influenced by factors like reverse causality. People might hesitate to seek assistance for simple tasks they can manage independently, but the value of assistance becomes evident when they face a substantial workload. To clarify these relationships, future research should consider confounding variables and mediator analysis. Our study offers valuable insights, underlining the necessity for further exploration to conclusively establish causation.

Future Research Directions

- Delve deeper into the factors behind reverse causality in the SUS and RTLX relationship.
- Investigate potential mediator variables to shed light on the observed correlation.
- Evaluate the influence of extraneous variables, including participant demographics, task complexity, and environmental conditions.
- Extend the study with more extensive and diverse datasets to bolster generalizability.
- Examine long-term trends in the SUS and RTLX scores relationship, unveiling potential changes over time.

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.
- Abadie, A. (2005). Causal Inference. Harvard University, Cambridge, Massachusetts, USA.
- DeCoster, J. (2005). Meta-Analysis. In Encyclopedia of Social Measurement (pp. 683–688). Elsevier.
- Pourhoseingholi, M. A., Baghestani, A. R., & Vahedi, M. (2012). How to control confounding effects by statistical analysis. Gastroenterol Hepatol Bed Bench, 5(2), 79-83.
- Zaniletti, I., Larson, D. R., Lewallen, D. G., Berry, D. J., & Maradit Kremers, H. (2022). How to Distinguish Correlation From Causation in Orthopaedic Research. The Journal of Arthroplasty, 38, 634-637.
- Devick, K. L., Zaniletti, I., Larson, D. R., Lewallen, D. G., Berry, D. J., &amp; Maradit Kremers, H. (2022). Avoiding systematic bias in orthopedics research through informed variable selection: A discussion of confounders, mediators, and Colliders. The Journal of Arthroplasty, 37(10), 1951–1955.
- Keith, K.A., Jensen, D.D., & O'Connor, B.T. (2020). Text and Causal Inference: A Review of Using Text to Remove Confounding from Causal Estimates. ArXiv, abs/2005.00649.
- Wunsch, G. (2007). Confounding and control. *Demographic Research, 16,* 97–120.
- Zwakman, D.S., Pal, D. & Arpnikanondt, C. Usability Evaluation of Artificial Intelligence-Based Voice Assistants: The Case of Amazon Alexa. SN COMPUT. SCI. 2, 28 (2021).