(NOTE THAT I'VE REMOVED THE WORK OF OTHERS SO AS NOT TO RISK MISREPRESENTING THEIR WORK AS MY OWN)

D4 - Analytic and Empirical Evaluation

Group 6

0.1 Results

0.1.1 Participant Data and Exclusions. 20 participants provided data for empirical evaluation. Of the 20, two were discounted from the data pool: one subject only had access to one of the systems, so further consideration of the differences between the two systems is impossible; the other subject did not have their data saved properly after the SUS survey, creating a gap in the data that prevented analysis. In total, 18 whole samples were collected.

Table 1. Performance Metrics Comparison of Three Systems for Event Recommendation

		System 1		System 2		System 3	
Metric	Description	Mean	SD	Mean	SD	Mean	SD
Total events displayed	Count of events shown on the screen.	38.9	6.7	38.2	3.4	161.2	66.7
Total events engaged with	Count of events the participant engaged with.	14.3	3.4	18.8	6.5	25.6	23.2
Total reading time	Aggregate time spent on event information.	62.6	49.5	87.1	52.7	91.7	77.0
Total events wished to attend	Count of events that the participant wished to attend.	9.7	2.4	12.8	4.1	20.3	22.1

0.1.2 Results Summary. Table 1 shows the summary statistics for the 4 metrics that were measured for each system. 18 samples were used to calculate the means and standard deviation for each system, for each metric.

System 3 displayed a significantly higher average number of events as well as a much wider range of measurements (M = 161.2, SD= 66.7) for each user session, compared to System 1 (M = 38.9, SD = 6.7), and System 2 (M=38.2, SD=3.4). The significantly higher average total events displayed is in line with a previous survey conducted about existing systems [1].

Across the remaining 3 metrics (Interest in event details, Total reading time, Interest indication), System 3 presented higher measurements on average (23.2, 77.0, 22.1) than Systems 1 and 2 (14.3, 62.6, 9.7, and 18.8, 87.1, 12.8 respectively), but also experienced more variability across those measurements, indicated by the differences in standard deviation. The higher means and variances can be attributed to more events being displayed to the user during the session.

The "Total reading time" metric showed the smallest difference in mean and variance between System 3 and Systems 1 and 2. Participants had 5 minutes to use each system, so reading time was constrained by the parameters of the experiment. This could have contributed to a smaller difference in variance compared to other metrics.

Table 2 shows the results after adjusting the average quantity of events the user engaged with and the ones they wished to attend for the average total number of events shown to the user. Note that this normalisation was only

Table 2. Recommended and Interesting Events Adjusted for Total Events.

	System 1		System 2		System 3	
Total-event ratios	Mean	SD	Mean	SD	Mean	SD
Ratio of Events Engaged With to Total Events	0.38	0.11	0.49	0.16	0.16	0.14
Ratio of Events wished to Attend to Total Events	0.26	0.08	0.34	0.10	0.13	0.12

Note: Ratios were calculated by dividing the metric value by the quantity of total events each user was served. The new sample population (n=18) was then used to calculate the mean and standard-deviation for each ratio.

possible under the assumption that the events the user would find attractive between systems was controlled as a result of the aforementioned "persona" system. This reveals more information about the efficiency and effectiveness of the systems display methods.

Between the systems, there is no significant pattern in the ranking of the rates of variability between the Ratio of Events Engaged With to Total Events (0.11, 0.16, 0.14) and Ratio of Events wished to Attend to Total Events (0.08, 0.10, 0.12). System 1 did show marginally smaller variability across the adjusted metrics, however.

After adjusting for total events, the amount of engagement per event curated by a system was significantly higher in System 1 (0.38) and even more so in System 2 (0.49) compared to System 3 (0.16). This is potentially because System 3 does not have a mechanism to recommend events to the users interests. Similarly, the number of events the user found interesting was ranked in order success: System 3 (0.13), System 1 (0.26), and finally System 2 (0.34).

0.1.3 Methods of Analysis. Statistical analysis was conducted to evaluate the differences between the “reference” and “comparison” systems. This analysis was important for assessing whether the systems had changed user engagement with societies, per the aim of the project.

Single-tailed, paired-sample t-testing was employed to assess significance of the differences in adjusted mean value. All combinations of systems were considered, for both adjusted metrics. The samples, whilst anonymous, were kept in 3-tuples to preserve the relationship between measurements across the three systems..

Cohen’s d value was used as a measure of effect size, and practical significance of the results in the real world. Whilst not typically negative, the sign of the Cohen’s d values was used to indicate the direction in which the difference is present. A Cohen’s d value equal to or larger than 0.8 is considered, in this context, large, indicating a high effect size and practical significance, whereas values in the range 0.5-0.8 are considered to have a medium effect size, and values between 0.3-0.5 are considered to have a small effect size.

A confidence level of 95% was used for all analyses to ensure that conclusions drawn from the results of analysis are likely to be representative of the wider user population. Assumptions of normality and paired observations were met, ensuring the appropriateness of the paired t-test for the data. A sample size of 18 was used after removing two erroneous samples from the pool of data - motivated by circumstances surrounding the collection of those samples that compromised validity.

The null hypothesis (H0) for analyses stated there would be no difference between the reference system and the comparison system using a given adjusted metric. The alternative hypothesis (H1) posited that there would be a statistically significant advantage, measured by the adjusted metric, for the reference system.

Table 3. Cohen's d Effect Size Matrix with Confidence Intervals: Ratio of Events Engaged With to Total Events

		System 1	System 2	System 3
System 1	Cohen's-d-value	-	**-0.77	***1.29
	p-value	-	.008	<.001
System 2	Cohen's-d-value	**0.77		***1.48
	p-value	.008	-	<.001

* p < .05, ** p < .01, *** p < .001
Note: Each cell in the matrix represents a comparison between two systems. The row indicates the "reference" system, while the column indicates the "comparison" system. A positive Cohen's-d value indicates that the reference system had a higher, adjusted-mean-value than the comparison system.

Table 4. Cohen's d Effect Size Matrix with Confidence Intervals: Ratio of Events wished to Attend to Total Events

		System 1	System 2	System 3
System 1	Cohen's-d-value	-	***-0.80	***1.38
	p-value	-	<.001	<.001
System 2	Cohen's-d-value	***0.80		***1.38
	p-value	<.001	-	<.001

* p < .05, ** p < .01, *** p < .001
Note: Each cell in the matrix represents a comparison between two systems. The row indicates the "reference" system, while the column indicates the "comparison" system. A positive Cohen's-d value indicates that the reference system had a higher, adjusted-mean-value than the comparison system.

0.1.4 Results of Analysis. The first adjusted metric table 3, the proportion of Good recommendations to Total Events Shown, indicates multiple statistically significant (p < .01) advantages. Table 4 shows similar patterns for the other adjusted metric, the Ratio of Interesting Events to Total Events, that are also highly statistically significant (p <.001):

- Both System 1 and System 2 have statistically significant advantages over System 3 for both adjusted metrics, with System 2 exhibiting the largest, significant, advantage (d = 1.48, 1.38, p <.001) relative to System 1 (d = 1.29, 1.38, p < .001). This implies, with a high level of confidence, that Systems 1 and 2 are both superior to System 3 at displaying events in a fashion that lead them to be more likely to be engaged with. H0 was rejected and the alternative hypothesis accepted.
- System 1 has a significant, negative, difference (d = -0.77) in the proportion of events served to the users that they perceived as engagement-worthy compared to System 2. This is a statistically significant conclusion (p = .008), and thus leads us to reject the null hypothesis and conclude that System 2 has better performance under the first metric. That is, for the events served to users on System 2, users were more likely to perceive them as being a Good Recommendation than if the same events were served on System 1.

  Similarly, the second metric (Table 3) shows that System 2 presents a statistically significant advantage (d = -0.80, p < .001) over System 1. That is, of the events served to the user they were more likely to engage with them if they were using System 2 as opposed to System 1.

  Both Systems 1 and 2 exhibited advantages in both adjusted metrics over System 3, with System 2 showing a relatively larger difference. This advantage is in itself statistically significant, as demonstrated by the direct comparison of System 1 to System 2 (d = 0.80, p < .001).

- System 1 has a significant, positive, difference (d = 1.38) when compared to System 3. Based on these results, the null hypothesis is rejected and one concludes that System 1 has a significant advantage over System 3. It should be noted that, on average, System 3 presented more events (table 1) than System 1, and System 1 presented events proportionally in a more engaging manner.

The same conclusions were drawn when analysing the opposite configuration of systems as the methods of analysis are symmetric up to swapping of reference and comparison systems.

Interestingly, as Systems 1 and 2 ultimately used the same recommendation algorithm and drew from the same pool of events, it's highly likely the changed feature present on System 2 - presenting more events at once - was solely responsible for influencing the increased engagement per event the participants reported.