HCI Evaluation II: Quantitative

Tags

- Quantitative Evaluation provides objective numerical data useful for statistical analysis.

In this lecture :

□ Questionares

□ NASA Task Load Index (NASA TLX)

□ System Usability Scale (SUS)

□ Statistical tests to determine if the perceived workload or system usability score has changed significantly

---

Questionares

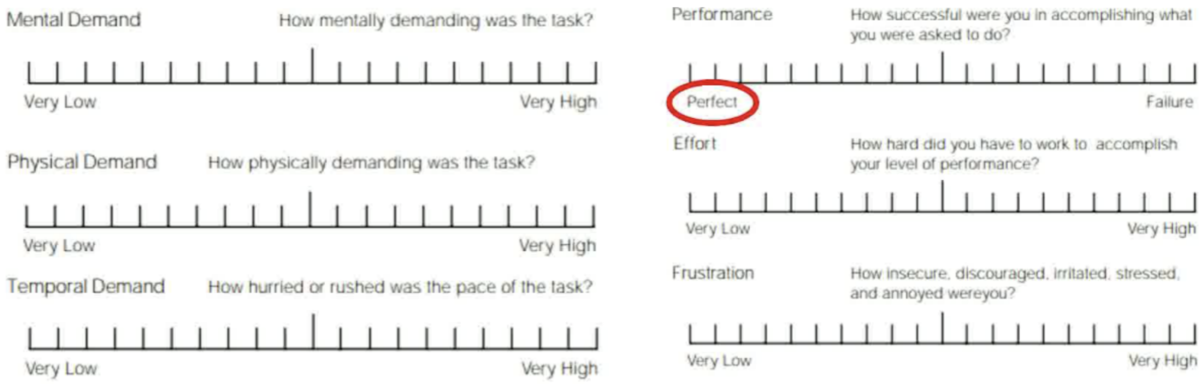
- Involve asking people to answer questions either on paper or digitally
- They can be used at scale with low resource requirements
- Generate a collection of demographics
- Evaluate designs and user requirements

NASA TLX

- The NASA Task Load Index is a questioner that estimates a user's perceived workload when using a system
- Workload is a complex construct but essentially means the amount of effort people have to exert both mentally and physically to use a system
- NASA TLX has over 8000 citations because it is viewed as the gold standard for **subjective workload**

The NASA TLX measures workload by looking at these various demands

- **Mental** : how much mental and perceptual activity was required
- **Physical :** how much physical activity was required
- **Temporal** : how much time pressure did the user feel due to the rate at which the tasks required completion
- **Performance :** How successful did the user feel when the accomplished the task
- **Effort :** How hard did the user have to work
- **Frustration** : how insecure did you feel.


Mental Demand
How mentally demanding was the task?
Very Low
Very High
Physical Demand
How physically demanding was the task?
Very Low
Very High
Temporal Demand
How hurried or rushed was the pace of the task?
Very Low
Very High
Performance
How successful were you in accomplishing what you were asked to do?
Perfect
Failure
Effort
How hard did you have to work to accomplish your level of performance?
Very Low
Very High
Frustration
How insecure, discouraged, irritated, stressed, and annoyed wereyou?
Very Low
Very High

There are two steps in scoring the questionare:

1. Identify how important each of the 6 demands was on the users perceived work load
2. Score the 6 demands on a scale

Relative weighting of the NASA TLX demands

- In this scoring method, (which is largely ignored nowadays, instead focusing on just the scores on each of the 6 demand scales) users are asked to compare which of the demands they thought was more relevant for the task they were asked to accomplish
- For example, for a game test, where the objective of the game is to stay alive, you could ask the users to compare the scales : Mental Demand vs Temporal Demand as being more relevant for the game, and because the aim of the game is just to survive with out any time limits most users would say mental demand was higher.
- There are a total of 15 paired combinations of the scales
- Each time a demand is selected as more important it receives a score of 1
- The total score is the weight of the dimension and a ranges for 0 to 5, (because you are comparing each dimension against 5 others)
- The max sum of the weights should be 15.

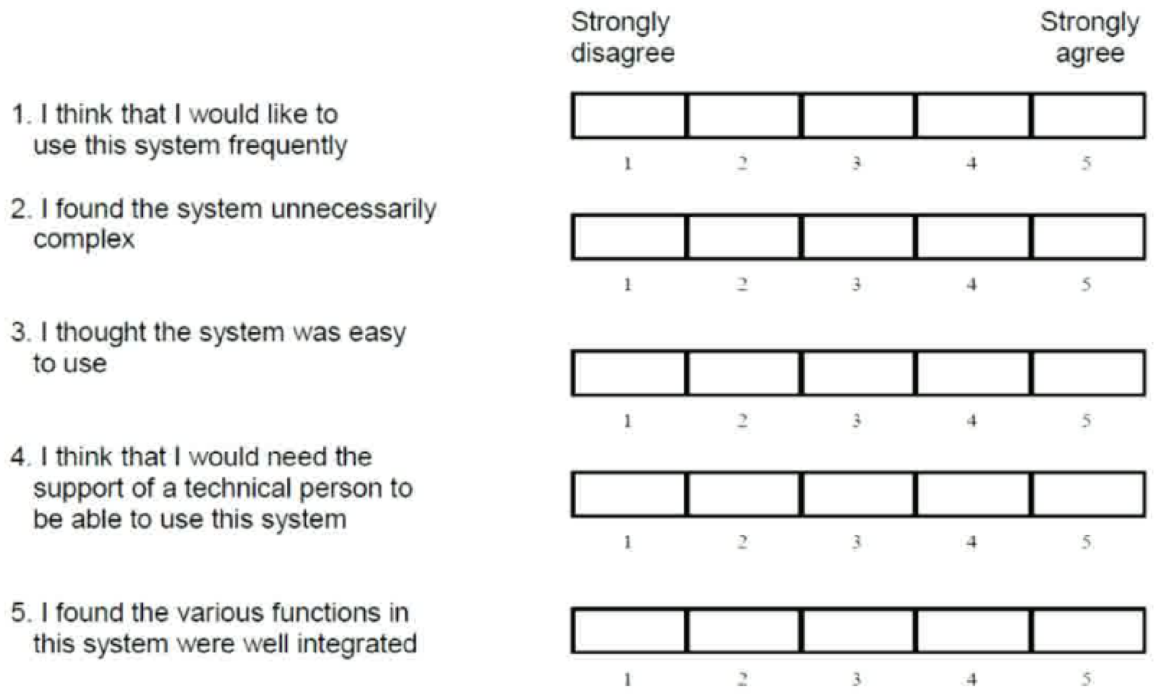
NASA TLX Rating the dimensions

- For each demand you will have a mark of 1 - 21
- If a user marks between the two ticks the right hand side of the tick is taken
- The score on a dimension is calculated as the tick number (1, 21) - 1 * 5
    - Example: If someone said that the mental demand of a task was 7, then the score would be : (7 - 1) * 5 = 30
- The total score is the sum of the 6 scores, which gives your a score between 0 - 100
- This is when you are not using the weighting (the first part of the score where each demand is weighed against the others) and hence is called a raw NASA TLX score

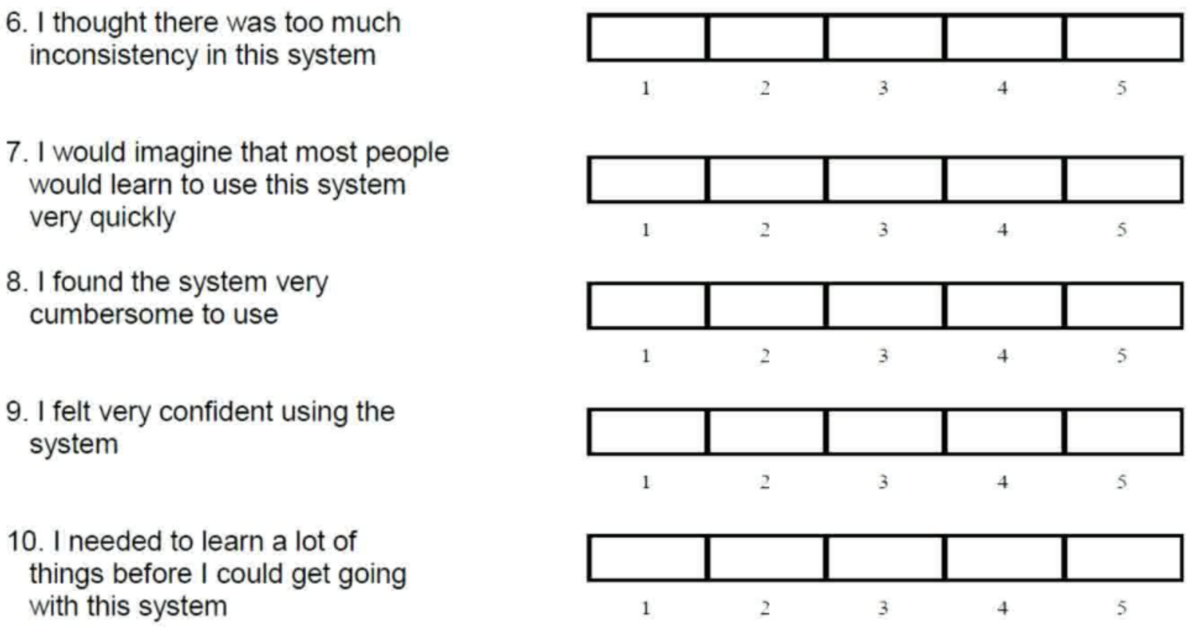
System Usability Scale

- This scale has been described as “quick and dirty”
- 10 questions with 5 responses are given to the user, the scaling is from Strongly agree to Strongly disagree

The questions can be whatever you want, but the SUS also provides questions of it own :



	Strongly disagree				Strongly agree
1. I think that I would like to use this system frequently	□ 1	□ 2	□ 3	□ 4	□ 5
2. I found the system unnecessarily complex	□ 1	□ 2	□ 3	□ 4	□ 5
3. I thought the system was easy to use	□ 1	□ 2	□ 3	□ 4	□ 5
4. I think that I would need the support of a technical person to be able to use this system	□ 1	□ 2	□ 3	□ 4	□ 5
5. I found the various functions in this system were well integrated	□ 1	□ 2	□ 3	□ 4	□ 5



6. I thought there was too much inconsistency in this system	□ 1	□ 2	□ 3	□ 4	□ 5
7. I would imagine that most people would learn to use this system very quickly	□ 1	□ 2	□ 3	□ 4	□ 5
8. I found the system very cumbersome to use	□ 1	□ 2	□ 3	□ 4	□ 5
9. I felt very confident using the system	□ 1	□ 2	□ 3	□ 4	□ 5
10. I needed to learn a lot of things before I could get going with this system	□ 1	□ 2	□ 3	□ 4	□ 5

- The question structure is designed where if the system was easy to use (positive responses) the odd numbered questions you would put “Strongly Agree”, but the even numbered questions you would put “Strongly disagree”
- Each item score contribution will range from 0 - 4
- For the Odd numbered questions the score contribution is the scale position -1
- For the even numbered questions the score is 5 - the scale position
  - example, for question 1 if the answer is 4, then the score is 3.
  - for question 2, if the answer is 2, then the answer is (5-2) = 3.
- The total score is then calculated by adding all the scores together, the max is 40 the min is 0
- This number is multiplied by 2.5 to obtain the overall score
- The total score is between 0 - 100
- If you have a SUS score **above 68** then your system has good usability (above average), anything below 68 is therefore below average
- A **key** difference between the NASA TLX and the SUS scoring system is that in the NASA TLX looking at the individual scores on each dimension is useful as

as evaluating the total score, whereas in the SUS system, it is advised only the total score is used

---

Statistical Testing

- Typically, you would test your system like software, game against a similar one, then preform a statistical analysis

Wilcoxon Test

- This test is ideal for small users 5 or less
- It is used when one user carries out two evaluations (like comparing two different difficulty levels of a game)

Preforming Statistical Test:

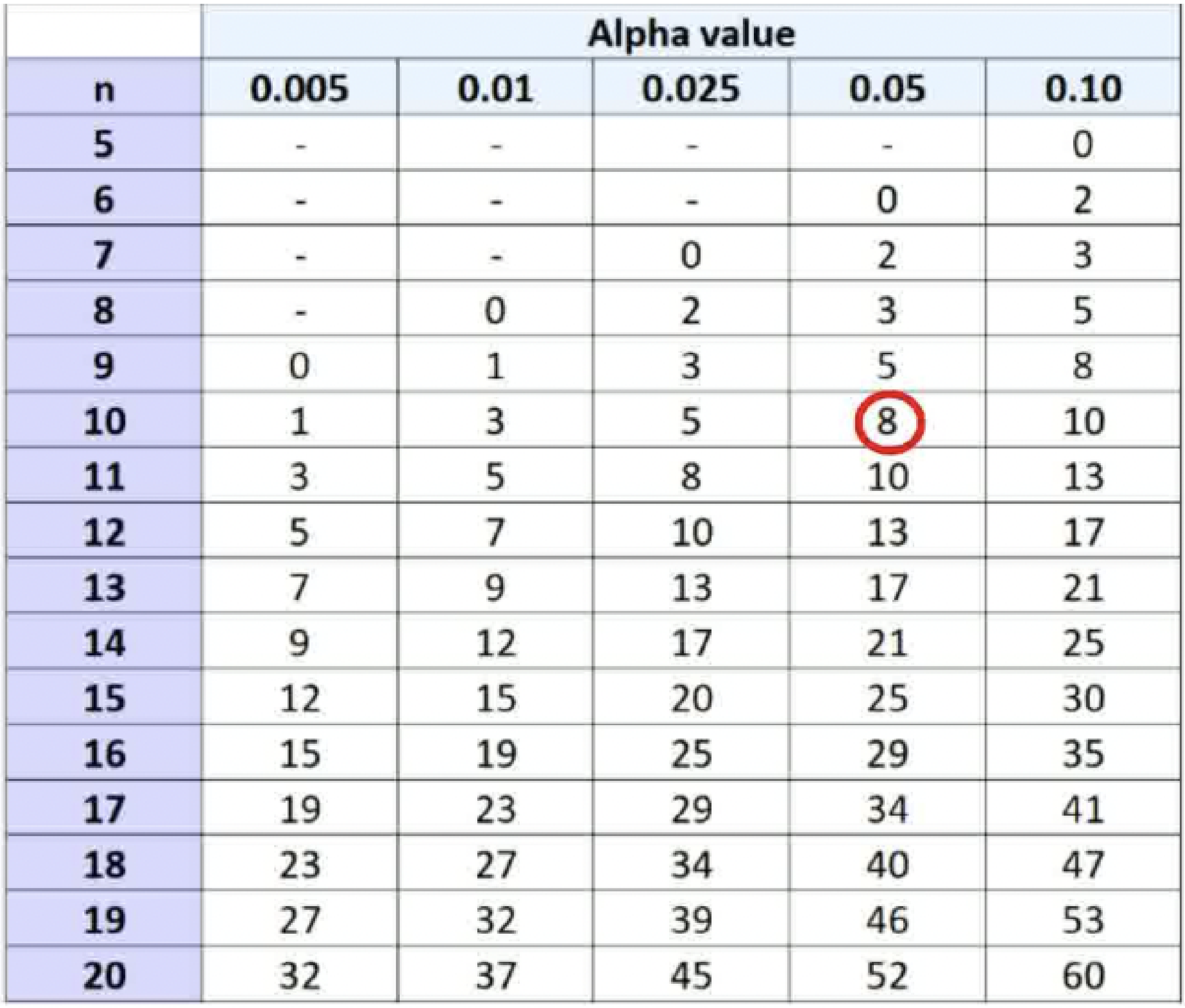
Example: Games testing on Easy and Hard Mode

1. Make a table where each row represents a users scores and two columns representing the easy and hard modes of the game

User	Easy Mode	Hard Mode
U1	25	67
U2	21	56
U3	18	43

- The Wilcoxon test will produce a **W value (W test statistic)** given N number of users.
- A significant result is a p value that is below 0.05. This means that there is a 95% probability that the results are due to a **real** difference and 5 % chance it is due to random chance

- The **W value** depends on the significance value you set and the number of users. Example if you have 10 users and a set a p value (Alpha value) of 0.05 then in order for there be a significant result, the corresponding **W value : must be below 8**



	Alpha value				
n	0.005	0.01	0.025	0.05	0.10
5	-	-	-	-	0
6	-	-	-	0	2
7	-	-	0	2	3
8	-	0	2	3	5
9	0	1	3	5	8
10	1	3	5	8	10
11	3	5	8	10	13
12	5	7	10	13	17
13	7	9	13	17	21
14	9	12	17	21	25
15	12	15	20	25	30
16	15	19	25	29	35
17	19	23	29	34	41
18	23	27	34	40	47
19	27	32	39	46	53
20	32	37	45	52	60

- The Wilcoxon test is useful for the same demographic of people comparing two different scenarios / systems, e.g. here the users are assumed to have the same level of gaming knowledge.

Mann-Whitney U test:

- If you want to test two different demographics of users such as testing experiences gamers vs non experienced gamers, then you would use the Mann Whitney U test.

How the Wilcoxon Signed-Rank Test Works

The Wilcoxon Signed-Rank Test operates under the following steps:

1. **Data Pairing**: Each subject or item in the study provides two scores, which often represent a "before" and "after" measurement (e.g., pre-test and post-test scores).
2. **Calculate Differences**: For each pair, calculate the difference. Exclude pairs with no change (where the difference is zero).
3. **Rank the Differences**: Rank the absolute values of the differences. The smallest difference gets a rank of 1, the next smallest a rank of 2, and so on up to the largest difference.
4. **Assign Signs to Ranks**: Assign signs to the ranks based on the signs of the differences. If the difference (post-test minus pre-test) is positive, the rank gets a positive sign; if the difference is negative, the rank gets a negative sign.
5. **Sum the Ranks**: Calculate the sum of the positive ranks (S+) and the sum of the negative ranks (S-).
6. **Test Statistic**: The test statistic for the Wilcoxon Signed-Rank Test is typically the smaller of the two sums (S+ and S-). This statistic is compared against a distribution of summed ranks for all possible combinations of positive and negative rank assignments under the null hypothesis of no effect.
7. **Decision Rule**: Based on the calculated statistic and the critical values from the Wilcoxon distribution (which depend on the sample size), decide whether to reject the null hypothesis. The null hypothesis usually states that the median difference between the two sets of scores is zero (i.e., no effect).

Example

Imagine you are testing a new teaching method, and you want to know if it significantly improves students' scores. You have test scores for 10 students before and after implementing the new method:

- **Before**: [82, 90, 78, 84, 70, 75, 88, 92, 80, 85]
- **After**: [85, 92, 80, 88, 75, 78, 90, 94, 82, 87]

The steps would be as follows:

1. Calculate the difference for each student (`After - Before`).
2. Rank the absolute differences.
3. Assign signs to each rank based on whether the difference was positive or negative.
4. Sum the ranks of the positive and negative differences.
5. Use the smaller sum to find the p-value or compare it against a critical value from the Wilcoxon Signed-Ranks table.

The decision to reject or not reject the null hypothesis will depend on the significance level you choose (commonly 0.05). Rejecting the null hypothesis suggests a significant difference in the median scores before and after the teaching method was applied, implying the method had an effect.

The Wilcoxon Signed-Rank Test is particularly useful because it does not assume that the differences are normally distributed, making it a robust test for non-normal data. It's widely used in situations where data is ordinal or when interval data fails tests for normality, providing a reliable method for assessing changes or effects in paired data.