8.HCI Evaluation p2

This lecture is about:

- Questionnaires
- NASA Task Load Index (NASA TLX)
- System Usability Scale (SUS)
- A statistical tests to determine if the perceived workload or system usabiity score has changed significantly

Questionnaires - defined

- Questionnaires involve asking people to answer questions either on paper or digitally e.g. on a webpage or app
- They can be used at scale with low resource requirements
- They generate a collection of demographic data and user opinions
- They can be used to evaluate designs and for understanding user requirements

Questionnaires are a type of survey used in HCI to gather data about users' attitudes, opinions, and preferences related to a product or system. They are often used to supplement other evaluation methods such as usability testing or heuristic evaluation.

Questionnaires can be administered in different formats, including paper-based, online, or through email. They can include a variety of question types such as multiple choice, Likert scales, open-ended questions, and ranking questions. The questions can be designed to collect data about different aspects of the user experience, such as ease of use, satisfaction, perceived usefulness, and trust.

When designing a questionnaire, it's important to consider the target audience and ensure the questions are clear, concise, and easy to understand. It's also important to avoid leading questions or bias, and to test the questionnaire with a small sample of users before administering it widely.

Once the questionnaire is administered, the data collected can be analyzed to identify patterns or trends in users' attitudes and opinions. This information can then be used to inform design decisions or improvements to the product or system.

Questionnaires - tips

- Ensure that you are asking a feasible number of questions (question fatigue is a thing)
- Watch out for leading questions e.g. “Why did you have difficulty with the navigation?”

引导性问题是一种提示或鼓励所需答案或回应的问题。它可能会提出一个特定的答案或包含回答问题的人不一定同意的假设。在给出的示例中，“您为什么在导航方面遇到困难？”，该问题已经假定此人在导航方面遇到困难，但情况可能不一定如此。相反，可以问一个更中立和开放性的问题，例如“你能描述一下你的导航体验吗？”。这允许人们提供他们自己的观点，而无需被引导到一个特定的答案。

- It is difficult to produce your own questionnaires
- It is best to use existing questionnaires that have been validated i.e. they measure what they claim to be measuring
- It's important to consider the language and tone of the questions to ensure that they are clear and easy to understand for the target audience.
- It's also helpful to have a mix of closed-ended (e.g. multiple choice, Likert scale) and open-ended questions to gather both quantitative and qualitative data.
- Piloting the questionnaire with a small group of participants before using it on a larger scale can help identify any issues with the questions or response options.

NASA TLX

NASA TLX (Task Load Index) is a questionnaire used to measure the subjective mental

workload experienced by an individual while performing a task. It was originally developed by NASA in the 1980s to evaluate the workload of pilots, but has since been applied to a wide range of tasks and industries. The questionnaire consists of six subscales: mental demand, physical demand, temporal demand, performance, effort, and frustration. Participants rate each subscale on a 0-100 scale, with higher scores indicating greater perceived workload. NASA TLX can provide valuable insights into the cognitive demands of a task and help identify areas that need improvement. NASA TLX（任务负荷指数）是一种问卷，用于衡量个人在执行任务时所经历的主观心理负荷。它最初由 NASA 在 1980 年代开发，用于评估飞行员的工作量，但后来被广泛应用于各种任务和行业。问卷由六个分量表组成：心理需求、身体需求、时间需求、表现、努力和挫折感。参与者在 0-100 范围内对每个子量表进行评分，分数越高表示感知到的工作量越大。NASA TLX 可以提供对任务认知需求的宝贵见解，并帮助确定需要改进的领域。

- The NASA Task Load Index (TLX) is a questionnaire that estimates a user's perceived workload when using a system. NASA任务负荷指数（TLX）是一份调查问卷，用于估计用户在使用一个系统时的感知工作量。
- Workload is a complex construct but essentially means the amount of effort people have to exert, both mentally and physically, to use a system. 工作量是一个复杂的结构，但基本上意味着人们在使用一个系统时，在精神上和身体上所要付出的努力。
- It was developed by Sandra Hart of NASA's human performance group and Lowell Staveland of San Jose University.它是由美国宇航局人类表现小组的Sandra Hart和圣何塞大学的Lowell Staveland开发的。
- The focus is on measuring the "immediate often unverbalized impressions that occur spontaneously" (Hart and Staveland, 1988). These are difficult or impossible to observe objectively.其重点是测量 "自发发生的直接的、往往是不言而喻的印象"（Hart和Staveland, 1988）。这些是很难或不可能客观地观察到的。
- Originally the NASA TLX questionnaire was developed for use in aviation but it's since been used in many different domains, including air traffic control, robotics, the automative industry, healthcare, website design and .other technology fields.最初，NASA的TLX调查表是为航空业开发的，但后来它被用于许多不同的领域，包括空中交通管制、机器人、自动化工业、医疗保健、网站设计和其他技术领域。

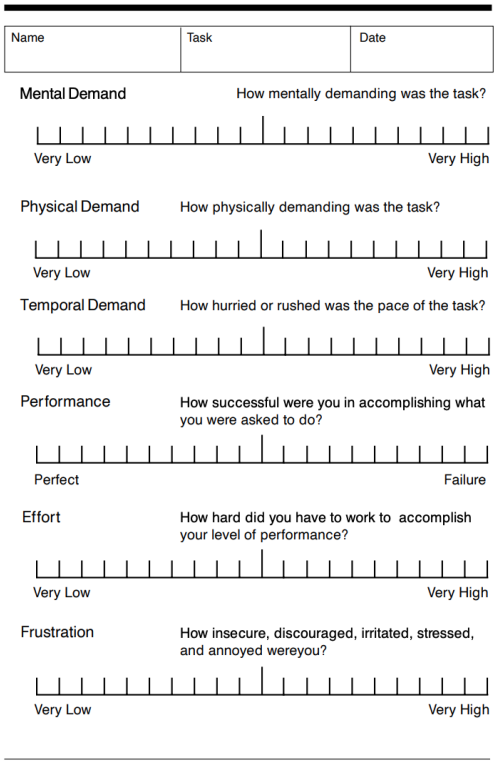
- Since it was introduced in 1988, it has had over 8000 citations. 自1988年推出以来，它已经有超过8000次的引用。
- It is viewed as the gold standard for measuring subjective workload.它被认为是测量主观工作量的黄金标准。
- Originally it was developed as a paper and pencil questionnaire but there are also free apps for iOS and Android
- The official website is here: https://humansystems.arc.nasa.gov/groups/TLX/index.php

This is paper and pen version:

Figure 8.6

NASA Task Load Index

Hart and Staveland's NASA Task Load Index (TLX) method assesses work load on five 7-point scales. Increments of high, medium and low estimates for each point result in 21 gradations on the scales.



The NASA TLX uses a multi-dimensional rating procedure that derives an overall workload score based on a weighted average of ratings on **six subscales**:

1. **Mental Demand**: This subscale measures how much mental and perceptual activity was required to perform the task. It takes into account factors such as the complexity of the task, the need for concentration and attention, and the level of mental effort required.
2. **Physical Demand**: This subscale measures how much physical activity was required to perform the task. It includes factors such as the amount of manual handling or physical effort required to carry out the task.
3. **Temporal Demand**: This subscale measures the level of time pressure felt by the user during the task. It takes into account factors such as the pace at which tasks occurred, the frequency and timing of deadlines, and the level of urgency associated with the task.
4. **Performance**: This subscale measures how successful the user felt in accomplishing the task. It takes into account factors such as how insecure, discouraged or irritated the user felt during the task, and how well they think they performed relative to their expectations.
5. **Effort**: This subscale measures how much effort (both mental and physical) was required to accomplish the user's level of performance. It takes into account factors such as the amount of concentration and attention required, the level of difficulty of the task, and the level of fatigue experienced during the task.
6. **Frustration**: This subscale measures how successfully the user felt they accomplished the task. It takes into account factors such as how well the user felt they understood the task, how satisfied they were with the outcome, and how well they felt they were able to manage the task.

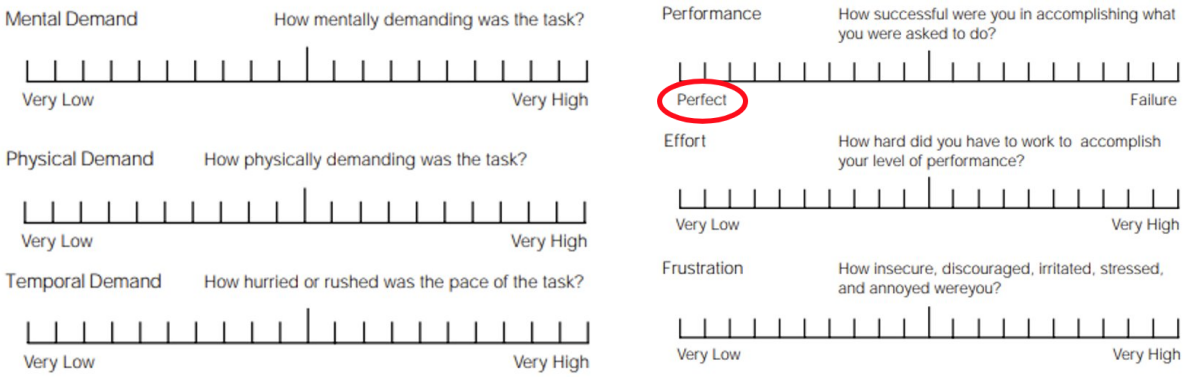
NASA TLX Scoring

NASA TLX scoring involves a weighting process to determine the relative contribution of each of the six subscales to the overall workload score. First, the participant is asked to allocate a total of 100 points among the six subscales based on how much they felt each subscale contributed to their overall workload. Then, the weights are multiplied by the subscale scores (ranging from 0 to 100) to give a weighted subscale score. The weighted subscale scores are added together to give an overall workload score ranging from 0 to 100. A higher score indicates a higher workload.

NASA TLX 评分涉及一个加权过程，以确定六个分量表中每一个分量表对总体工作量分数的相对贡献。首先，要求参与者根据他们认为每个分量表对他们整体工作量的贡献程度，在六个分量表中分配总共 100 分。然后，将权重乘以子量表分数（范围从 0 到 100）以给出加权子量表分数。将加权子量表分数相加得到 0 到 100 之间的总工作量分数。分数越高表示工作量越大。

It is important to note that the NASA TLX scoring process is subjective and can vary depending on the individual's interpretation and weighting of the subscales. Therefore, it is recommended that multiple participants evaluate the same task and that the scores are averaged to get a more reliable overall workload score.

重要的是要注意，NASA TLX 评分过程是主观的，可能会因个人对分量表的解释和权重而有所不同。因此，建议多个参与者评估同一个任务，并将分数平均以获得更可靠的整体工作量分数。



- Users answer the NASA TLX after they have completed a task. This is necessary as asking them to complete it during task is typically not possible. However, it may mean that users forget details of the perceived workload.
- The questionnaire is scored in a two step process:

1. Identifying the relative importance of the 6 dimensions on a user’s perceived workload
2. Rating each of the 6 dimensions on a scale

NASA TLX Relative weighting of dimensions

The relative weighting of dimensions in the NASA TLX can vary depending on the specific task being evaluated and the preferences of the user. In general, however,

Mental Demand and Physical Demand tend to be the most heavily weighted dimensions, as they typically require the most cognitive and physical effort from the user. Temporal Demand, Performance, and Effort are also important dimensions to consider, as they can greatly impact the user's workload and satisfaction with the task. Frustration is often the least heavily weighted dimension, as it is typically a result of other dimensions being high (e.g. mental demand leading to frustration). However, it is still important to consider as it can have a significant impact on the user's overall experience. NASA TLX 中维度的相对权重可能会有所不同，具体取决于所评估的具体任务和用户的偏好。然而，一般来说，心理需求和身体需求往往是权重最大的维度，因为它们通常需要用户付出最多的认知和体力。临时需求、性能和工作量也是需要考虑的重要维度，因为它们会极大地影响用户的工作量和对任务的满意度。挫折感通常是权重最低的维度，因为它通常是其他维度高的结果（例如导致挫败感的精神需求）。但是，考虑这一点仍然很重要，因为它会对用户的整体体验产生重大影响。

- A user reflects on the task they've been asked to perform and is shown each paired combination of the six dimensions to decide which is more related to their personal definition of workload as related to the task. 用户对他们被要求执行的任务进行思考，并向他们展示六个维度的每一个配对组合，以决定哪一个与他们个人对任务相关的工作量定义更相关。
- This means a user considers 15 paired comparisons. For example, they need to decide whether Performance or Frustration "represents the more important contributor to the workload for the specific task you recently performed." 这意味着用户要考虑15个成对的比较。例如，他们需要决定是绩效还是挫败感 "代表了你最近执行的具体任务的更重要的工作量"。
- Each time a dimension is selected as more important it receives a score of 1. The total score is the weight of the dimension and ranges from 0 to 5.每当一个维度被选为更重要的时候，它就会得到1分。总分是该维度的权重，范围从0到5。
- The sum of the weights should be 15. 权重的总和应该是15。
- The relative weighting of the six dimensions is often not measured or used.六个维度的相对权重通常不被衡量或使用。
- This makes the NASA TLX simpler to administer. 这使得NASA的TLX在管理上更加简单。
- Several studies have compared raw TLX scores to weighted TLX scores and have found mixed results (some showing better sensitivity when removing weights, others

showing no difference, and others showing less sensitivity).一些研究比较了原始TLX分数和加权TLX分数，发现结果不一（有些显示去除权重后的敏感性更好，有些显示没有区别，有些显示敏感性更低）。

- When the dimensions are not rated the method is called the ‘raw TLX score’当维度没有被评级时，该方法被称为 "原始TLX得分"。

NASA TLX Rating the dimensions

- Users mark their score on each of the six dimensions. 用户在六个维度的每一个维度上标记他们的分数。
- Each dimension consists of a line with 21 equally spaced tick marks, which divide the line from 0 to 100 in increments of 5. If a user marks between two ticks then the value of the right tick is used.每个维度由一条线组成，线上有21个等距的刻度线，这些刻度线从0到100，增量为5。如果用户在两个刻度线之间做标记，则使用右边的刻度线的值。
- The score on a dimension is calculated as the tick number (1 – 21) – 1 multiplied by 5. 一个维度上的分数是以刻度线的数量（1-21）-1乘以5计算的。
- For example, the images show the rating on a paper questionnaire (top) and on a mobile app (bottom)例如，图片显示了纸质问卷（顶部）和移动应用程序（底部）上的评分。
- The fifth tick mark is selected, so the rating score is: **(5 – 1) * 5 = 20** 第五个勾号被选中，所以评级分数是：（5-1）*5=20

NASA TLX What do the scores tell us?

- If the weights are used then the individual ratings on each of the dimensions are multiplied by their respective weights, summed and divided by 15, resulting in an aggregate perceived workload score for a task ranging from 0 – 100. 如果使用了权重，那么每个维度上的单独评分都要乘以各自的权重，然后相加再除以15，从而得出一项任务的总体感知工作量分数，范围是0-100。
- If the weights are not used then the individual ratings on each of the dimensions can be summed and divided by 6, resulting in an aggregate perceived workload score ranging from 0 – 100. 如果不使用权重，那么每个维度上的个人评级可以被加总并除以6，从而得到一个总的感知工作量分数，范围为0-100。
- The individual ratings on the 6 dimensions also give some insight in to where the workload is coming from. This can be helpful for developers hoping to improve their design. 6个维度上的单独评分也能让人了解到工作量的来源。这对希望改进设计的开发者来说是有帮助的。

In the NASA TLX, the weights assigned to each dimension are used to calculate an aggregate perceived workload score for a task. If the weights are used, the individual ratings on each dimension are multiplied by their respective weights, summed, and then divided by 15, which results in an aggregate score ranging from 0-100. If the weights are not used, the individual ratings on each of the dimensions are summed and then divided by 6, resulting in an aggregate score ranging from 0-100. These scores help researchers and designers understand the overall perceived workload of a task and identify the specific dimensions that are contributing the most to the perceived workload. By analyzing the individual ratings on each of the dimensions, designers can gain insights into where the workload is coming from and make design improvements accordingly.

NASA TLX中，分配给每个维度的权重被用来计算一项任务的总体感知工作量分数。如果使用权重，每个维度上的单独评分乘以各自的权重，加总后再除以15，得出0-100的总分。如果不使用权重，则每个维度上的单独评分相加，然后除以6，得出0-100分的总分。这些分数可以帮助研究人员和设计人员了解一项任务的整体感知工作量，并确定对感知工作量贡献最大的具体维度。通过分析每个维度的单独评分，设计师可以深入了解工作量的来源，并据此进行设计改进。

Let's say we have a task that involves a user performing a series of math problems on a computer. After completing the task, the user is asked to rate their perceived workload using the NASA TLX questionnaire.

For each of the six dimensions (mental demand, physical demand, temporal demand, performance, effort, and frustration), the user is asked to rate the level of workload on a scale from 0 to 100. Let's say the user's ratings were as follows:

- Mental Demand: 70
- Physical Demand: 10
- Temporal Demand: 20
- Performance: 50
- Effort: 60
- Frustration: 40

If we use the weights, the individual ratings are multiplied by their respective weights (which were determined in the paired comparison exercise), summed, and divided by 15 to get an aggregate perceived workload score. Let's say the weights were:

- Mental Demand: 5
- Physical Demand: 1
- Temporal Demand: 2
- Performance: 3
- Effort: 4
- Frustration: 0

Then, the aggregate perceived workload score would be:

(70*5* + *10*1 + 20*2* + *50*3 + 60*4* + *40*0) / 15 = 46.67

If we don't use the weights, we simply sum up the individual ratings and divide by 6 to get the aggregate perceived workload score. In this case, it would be:

(70 + 10 + 20 + 50 + 60 + 40) / 6 = 41.67

So the weighted score is slightly higher than the unweighted score in this example. By looking at the individual ratings for each dimension, we can see that the user found the

task to be mentally demanding and requiring significant effort, but not very physically demanding or frustrating. This information can be helpful for developers to improve the design of the task and reduce perceived workload in future iterations.

How do we know if we want to use weight or not?

If the goal is to have a more nuanced understanding of the perceived workload on each of the six dimensions, then using weights would be appropriate. This would allow for more detailed analysis and potentially reveal specific areas where improvements could be made.如果目标是对六个维度中每个维度的感知工作量有一个更细微的了解，那么使用权重将是合适的。这将允许进行更详细的分析，并有可能揭示出可以进行改进的具体领域。

However, if the goal is to simply get an overall score of perceived workload without delving into the individual dimensions, then using weights may not be necessary. In this case, the simpler calculation of summing and dividing by 6 would suffice.然而，如果目标只是为了得到一个感知到的工作量的总分，而不深入研究各个维度，那么使用权重可能就没有必要了。在这种情况下，用更简单的计算方法，即相加再除以6就足够了

Ultimately, the decision to use weights or not should be based on the research question, the goals of the evaluation, and the desired level of detail and specificity in the results.最终，是否使用权重的决定应基于研究问题、评估目标以及结果中所期望的详细和具体程度。

NASA TLX Validity

- Hart and Staveland validated that the sub-scales measure different sources of workload.
- Subsequent independent studies have also found that the NASA TLX is a valid measure of subjective workload (Rubio et al, 2004; Xiao et al, 2005).

Advantages of NASA TLX:

- It provides a comprehensive and detailed understanding of different sources of workload.它提供了对不同来源的工作量的全面和详细的了解。
- It is a flexible and adaptable tool that can be used in a variety of settings.它是一个灵活的、适应性强的工具，可以在各种场合使用。

- It provides a quantitative measure of workload, which can be useful for comparisons between different tasks, systems, or designs.它提供了对工作量的定量测量，这对不同任务、系统或设计之间的比较很有用。
- It can provide valuable insights into specific aspects of a task or system that may need improvement.它可以为一项任务或系统中可能需要改进的具体方面提供有价值的见解。
- It is relatively easy to administer and can be completed quickly.它相对容易管理，可以快速完成

Disadvantages of NASA TLX:

- It is a subjective measure, and may be influenced by individual biases and experiences.它是一个主观的衡量标准，可能会受到个人偏见和经验的影响。
- It may not be appropriate for all types of tasks or systems, and may not capture all aspects of workload or usability.它不一定适合所有类型的任务或系统，也不一定能反映工作量或可用性的所有方面。
- It may not provide clear guidance for designers or developers on how to improve the usability or workload of a system.它可能无法为设计者或开发者提供明确的指导，说明如何改善系统的可用性或工作量。
- It may require additional training or expertise to administer and interpret the results accurately.它可能需要额外的培训或专业知识来管理和准确解释结果。

System Usability Survey (SUS)

The System Usability Scale (SUS) is a widely used questionnaire-based tool for evaluating the perceived usability of a system. It was developed by John Brooke in 1986 and consists of 10 items, each with a five-point Likert scale response (ranging from Strongly agree to Strongly disagree).

- The System Usability Scale (SUS) provides a “quick and dirty”, reliable tool for measuring usability.
- It was created by John Brooke in 1986.
- It consists of a 10 item questionnaire with five response options for respondents; from Strongly agree to Strongly disagree.

- It enables you to evaluate a wide variety of products and services, including hardware, software, mobile devices, websites and applications.

The items of the SUS are:

1. I think that I would like to use this system frequently.
2. I found the system unnecessarily complex.
3. I thought the system was easy to use.
4. I think that I would need the support of a technical person to be able to use this system.
5. I found the various functions in this system were well integrated.
6. I thought there was too much inconsistency in this system.
7. I would imagine that most people would learn to use this system very quickly.
8. I found the system very cumbersome to use.
9. I felt very confident using the system.
10. I needed to learn a lot of things before I could get going with this system.

The SUS score is calculated by converting the responses to each item into a score between 0 and 4, then adding up the scores and multiplying by 2.5. The resulting score is a number between 0 and 100, with higher scores indicating greater perceived usability. 总数相加*2.5

The SUS has been found to be a reliable and valid tool for measuring perceived usability across a wide range of systems and contexts. It is relatively quick and easy to administer, making it a popular choice for usability testing and evaluation.

Benefits of using a SUS

SUS has become an industry standard, with references in over 1300 articles and publications. The noted benefits of using SUS include that it:

- Is a very easy scale to administer to participants
- Can be used on small sample sizes with reliable results
- Is valid – it can effectively differentiate between usable and unusable systems

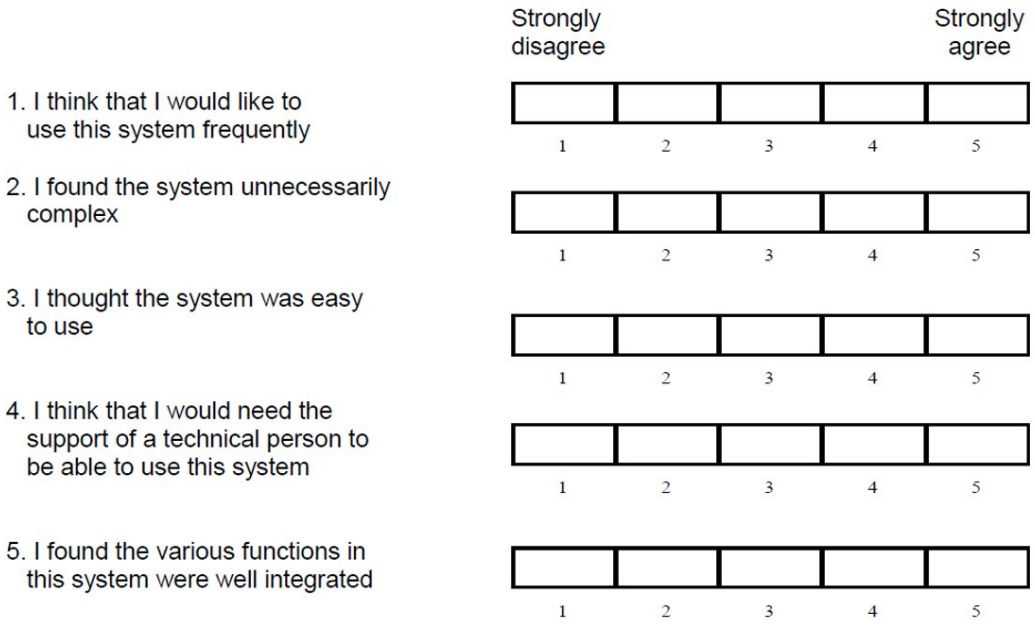
Considerations when using a SUS

If you are considering using a SUS, keep the following in mind:

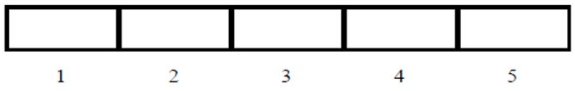
- The scoring system is somewhat complex
- There is a temptation, when you look at the scores, since they are on a scale of 0-100, to interpret them as percentages, they are not
- The best way to interpret your results involves "normalizing" the scores to produce a percentile ranking
- SUS is not diagnostic - its use is in classifying the ease of use of the site, application or environment being tested

System Usability Survey (SUS) - scale

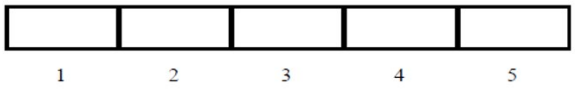
When an SUS is used, participants are asked to score the 10 items with one of five responses that range from Strongly Agree to Strongly disagree i.e. using Likert scales



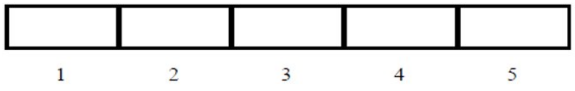
6. I thought there was too much inconsistency in this system



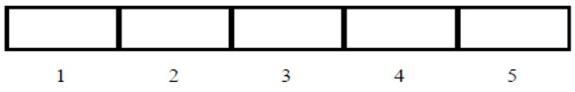
7. I would imagine that most people would learn to use this system very quickly



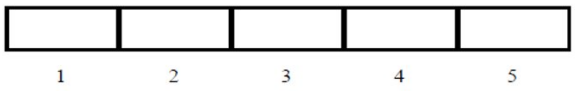
8. I found the system very cumbersome to use



9. I felt very confident using the system



10. I needed to learn a lot of things before I could get going with this system



System Usability Survey (SUS) – scoring

- The SUS is given to users when they have completed using the system which is being evaluate d当用户使用完被评估的系统后，会将SUS交给他们。
- They score each of the 10 items by marking one of the five boxes. 他们通过标记五个方框中的一个来给10个项目中的每一个打分。
- The SUS yields a single number representing a composite measure of the overall usability of the system being studied. Note that scores for individual items are not meaningful on their own. SUS产生一个单一的数字，代表被研究系统的整体可用性的综合测量。请注意，单个项目的分数本身是没有意义的。
- To calculate the SUS score, first sum the score contributions from each item. Each item's score contribution will range from 0 to 4. 要计算SUS分数，首先要将每个项目的分数贡献相加。每个项目的得分贡献将在0到4之间。
- For items 1,3,5,7,and 9 (the odd numbered items) the score contribution is the scale position minus 1. For items 2,4,6,8 and 10 (the even numbered items) the contribution is 5 minus the scale position. 对于项目1、3、5、7和9（奇数项目），其分数贡献是比例尺位置减去1。对于项目2,4,6,8和10（偶数项目）的贡献是5减去比例位置。(下面有例子展示)
- Multiply the sum of the scores by 2.5 to obtain the overall score. 分数之和乘以2.5，得到总分。

- SUS scores have a range of 0 to 100. SUS的分数范围为0至100。
- Based on research, a SUS score above a 68 would be considered above average and anything below 68 is below average. 根据研究，SUS分数高于68分，就被认为是高于平均水平，低于68分就是低于平均水平。

The scoring of the System Usability Scale (SUS) questionnaire involves assigning a score from 0 to 4 to each of the 10 items. For the odd-numbered items (1, 3, 5, 7, and 9), the score contribution is equal to the scale position (0-4) minus 1. For the even-numbered items (2, 4, 6, 8, and 10), the score contribution is equal to 5 minus the scale position (0-4).

This means that for the odd-numbered items, the scores are:

- 0 = 1 (原有的位置-1)
- 1 = 2
- 2 = 3
- 3 = 4
- 4 = 5

For the even-numbered items, the scores are:

- 0 = 5 (5-原有的位置)
- 1 = 4
- 2 = 3
- 3 = 2
- 4 = 1

This scoring system is used to create a final SUS score ranging from 0 to 100, with higher scores indicating better usability.

Advantages of System Usability Survey (SUS) include:

1. It is a quick and easy way to assess the overall usability of a system.
2. It provides a standardized measure of usability, allowing for comparisons across different systems.
3. It is a reliable measure, with high internal consistency and test-retest reliability.
4. It is a widely used and well-established measure of usability, with a large body of research supporting its validity.

Disadvantages of System Usability Survey (SUS) include:

1. It provides a general measure of usability, but does not provide specific information about the sources of usability problems.
2. It may not be sensitive enough to detect small differences in usability between systems.
3. It is a self-report measure, and users may not always accurately report their experiences or may be influenced by factors outside of the system being tested.
4. It does not provide guidance on how to improve usability, but rather identifies areas where improvement is needed.

Statistical testing

Statistical testing is a process of analyzing data and determining whether the results obtained are statistically significant or just due to chance. In the context of Human-Computer Interaction (HCI), statistical testing is often used to compare the performance of different interfaces or to assess the impact of design changes on user behavior.统计测试是一个分析数据的过程，并确定所得到的结果是有统计学意义的还是仅仅由于偶然性。在人机交互（HCI）的背景下，统计测试经常被用来比较不同界面的性能或评估设计变化对用户行为的影响。

Statistical testing involves defining a null hypothesis, which assumes that there is no significant difference between the groups being compared, and an alternative hypothesis, which assumes that there is a significant difference. Then, a statistical test is applied to the data to determine the likelihood of observing the data if the null hypothesis were true. If the likelihood is very low, typically less than 5%, the null hypothesis is rejected, and the alternative hypothesis is accepted.统计测试包括定义一个无效假设，即假设被比较的群体之间没有明显的差异，以及一个替代假设，即假设存在

明显的差异。然后，对数据进行统计测试，以确定如果无效假设是真的话，观察到数据的可能性。如果可能性很低，通常低于5%，则拒绝无效假设，并接受替代假设。

Some common statistical tests used in HCI include t-tests, ANOVA, and regression analysis. These tests can help HCI researchers and designers identify significant differences in performance, user satisfaction, or other user metrics between different interfaces or design variations. However, statistical testing requires careful consideration of sample size, data distribution, and other factors to ensure the results obtained are reliable and valid.一些在人机交互中常用的统计测试包括t检验、方差分析和回归分析。这些测试可以帮助人机交互研究者和设计者识别不同界面或设计变化之间在性能、用户满意度或其他用户指标方面的显著差异。然而，统计测试需要仔细考虑样本量、数据分布和其他因素，以确保获得的结果是可靠和有效的。

- You might get a user to rate the SUS of two different designs and want to know if one design is significantly better than the other 你可能会让用户对两种不同设计的SUS进行评价，并想知道一种设计是否明显优于另一种设计。
- Similarly, you might want to know if two levels of difficulty in your game are significantly different so you get a user to rate the workload of both levels 同样地，你可能想知道你的游戏中的两个难度级别是否有明显的不同，所以你让用户对两个级别的工作量进行评价
- To determine whether the differences in scores are significantly different we can use a statistical test为了确定分数的差异是否有明显的不同，我们可以使用统计测试
- There are many statistical tests but I am going to show you one that will be useful for your project 有很多统计测试，但我将向你展示一个对你的项目有用的统计测试
- It is the Wilcoxon Signed Rank Test and it is ideal for analysing data from Likert and other scales e.g. the NASA TLX and SUS这是Wilcoxon Signed Rank Test，它是分析Likert和其他量表数据的理想工具，例如NASA的TLX和SUS。
- It is used when one user carries out two evaluations e.g. rates the workload of your game at two different difficulty levels 当一个用户进行两个评价时，例如在两个不同的难度下对你的游戏的工作量进行评价时，它就被使用。
- It is a good test when you have small numbers of users – the minimum is 5; however, it's better at identifying significant differences when you have larger numbers of users 当你有少量的用户时，这是一个很好的测试--最少是5个；然而，当你有大量的用户时，它能更好地识别显著的差异。

- Make a table where each row represents a user's scores and each column a separate evaluation score.
- I've shown the results of three users evaluating the workload of a game at two difficulty levels using the NASA TLX.
- You need a minimum of 5 and ideally more

User ID	Workload level 1	Workload level 2
U1	25	67
U2	32	56
U3	18	43

- Enter the data into the online calculator: https://www.statology.org/wilcoxon-signed-rank-test-calculator/
- Look up the calculated W test statistic in the table of critical values在临界值表中查找计算出的W检验统计量
- To do this you need to know N, which is the number of users, and the significance level, which we will set at 0.05要做到这一点，你需要知道N，也就是用户的数量，以及显著性水平，我们将其设定为0.05
- This means that if a significant difference is found then it is 95% certain that this is a real difference rather than due to randomness这意味着，如果发现有显著性差异，那么95%的人可以肯定这是一个真正的差异，而不是由于随机性造成的

- We use an alpha value aka significance level of 0.05 我们使用0.05的α值（又称显著性水平）
- We find the row that corresponds to our number of users aka n 我们找到与我们的用户数（又称n）相对应的行

- If we have 10 users then the W test statistic generated by the online calculator needs to be less than 8 otherwise there is no significant difference 如果我们有10个用户，那么由在线计算器生成的W测试统计量需要小于8，否则就没有显著性差异。

	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

- If we are comparing two sets of values generated by two different groups e.g. experienced gamers and novice gamers then we use a different test to see if they are significantly different如果我们比较两组由两个不同群体产生的数值，例如有经验的玩家和新手玩家，那么我们使用不同的测试，看他们是否有显著差异
- This is known as the Mann-Whitney U test. There is also an online calculator and you can read about the test here:

https://www.statology.org/mann-whitney-u-test/

The Mann-Whitney U test is a non-parametric test used to compare two independent groups. It is also known as the Wilcoxon rank-sum test. The test is used to determine if there is a significant difference between the medians of two groups.

In the Mann-Whitney U test, the data from the two groups are combined and ranked from smallest to largest, and the rank sum for each group is calculated. The test statistic

U is calculated based on the rank sums and is used to determine the p-value. The null hypothesis is that there is no difference between the medians of the two groups.

The Mann-Whitney U test is appropriate for data that are ordinal, interval, or ratio, but not necessarily normally distributed. It is commonly used in behavioral and social sciences when sample sizes are small, and the data may not meet the assumptions of parametric tests.

In summary, the Mann-Whitney U test is a non-parametric test used to compare two independent groups, and it determines if there is a significant difference between the medians of the two groups.

Some quizes:

1. What does NASA TLX stand for?
   a) Time Load Index
   b) Task Level Experience
   c) Time Limited Experience
   d) Task Load Index
2. How many subscales does NASA TLX have?
   a) 4
   b) 5
   c) 6
   d) 7
3. What is the purpose of SUS?
   a) To measure the cognitive workload of users
   b) To measure the learnability of a system
   c) To measure the satisfaction of users with a system
   d) To measure the efficiency of a system
4. How many items are there in the SUS questionnaire?
   a) 5
   b) 8

c) 10
d) 12

5. In SUS, what is the scoring method for even numbered items?
   a) Scale position minus 1
   b) 5 minus the scale position
   c) Scale position plus 1
   d) 10 minus the scale position

6. What is the maximum score a system can receive on SUS?
   a) 50
   b) 75
   c) 100
   d) 125

7. What statistical test is commonly used to analyze SUS data?
   a) T-test
   b) ANOVA
   c) Mann-Whitney U test
   d) Chi-square test

8. Which of the following is not a subscale of NASA TLX?
   a) Performance
   b) Mental Demand
   c) Physical Demand
   d) Satisfaction

9. What is the purpose of NASA TLX?
   a) To measure the cognitive workload of users
   b) To measure the learnability of a system
   c) To measure the satisfaction of users with a system
   d) To measure the efficiency of a system

10. How is the total score calculated in NASA TLX when weights are used?
    a) By multiplying the sum of the ratings by 15
    b) By multiplying each rating by its respective weight, summing them, and dividing by 15
    c) By summing the ratings and dividing by 6
    d) By summing the ratings and dividing by 10

Answer:

1. d
2. c
3. c
4. c
5. b
6. c
7. c
8. d
9. a
10. b

More complicated

1. Which statistical test is used to compare two sets of paired data in usability testing?
a) T-test
b) ANOVA
c) Wilcoxon signed-rank test
d) Mann-Whitney U test

2. What is the purpose of A/B testing in usability evaluation?
a) To compare the performance of two different interfaces or designs
b) To measure the cognitive workload of users
c) To measure user satisfaction with a system
d) To determine the learnability of a system

3. Which of the following is not a type of usability test?
a) Heuristic evaluation
b) Think-aloud protocol

c) Expert review
d) Case study analysis

4. Which of the following is not a factor to consider when recruiting participants for usability testing?
a) Age
b) Gender
c) Education level
d) Political affiliation

5. What is the difference between quantitative and qualitative data in usability testing?
a) Quantitative data is numerical and can be measured objectively, while qualitative data is descriptive and subjective.
b) Qualitative data is numerical and can be measured objectively, while quantitative data is descriptive and subjective.
c) Quantitative data measures user satisfaction, while qualitative data measures performance.
d) Qualitative data measures user satisfaction, while quantitative data measures cognitive workload.

6. Which of the following is a method of analyzing qualitative data in usability testing?
a) Regression analysis
b) Content analysis
c) Factor analysis
d) ANOVA

7. What is the purpose of a task scenario in usability testing?
a) To give participants a clear understanding of the task they need to perform
b) To control for extraneous variables in the testing environment
c) To measure user satisfaction with a system
d) To measure the cognitive workload of users

8. What is the purpose of a debriefing session in usability testing?
a) To provide participants with feedback on their performance

b) To obtain participants' personal information
c) To measure user satisfaction with a system
d) To obtain participants' informed consent for the study

9. Which of the following is a limitation of heuristic evaluation?
a) It is expensive and time-consuming to carry out
b) It can only be used with finished products, not prototypes
c) It relies on the expertise of evaluators, who may have biases or different interpretations of heuristics
d) It requires a large sample size to obtain reliable results

10. Which of the following is a potential source of bias in usability testing?
a) The Hawthorne effect
b) Regression to the mean
c) Sampling bias
d) Confirmation bias

Answer:

1. c
2. a
3. d
4. d
5. a
6. b
7. a
8. a
9. c
10. c