HCI Evaluation - II

Questionnaires- 问卷调查表-定义

- 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 大规模低资源使用
- They generate a collection of demographic data and user opinions生成一系列人口统计数据和用户意见

They can be used to **evaluate designs** and for **understanding user requirements**

Questionnaires - tips **问卷调查表-技巧**

- Ensure that you are asking a **feasible number** of questions (question fatigue is a thing)数量合适
- Watch out for **leading questions** e.g. "Why did you have difficulty with the navigation?"避免引导性问题
- 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用验证过的问卷表

NASA TLX任务负荷指数

The NASA Task Load Index (TLX) is a questionnaire 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. 概念

工作负荷是一个复杂的概念，但基本上意味着人们在使用系统时需要付出的精力和努力，无论是精神上还是身体上。

It was developed by Sandra Hart of NASA's human performance group and Lowell Staveland of San Jose University.

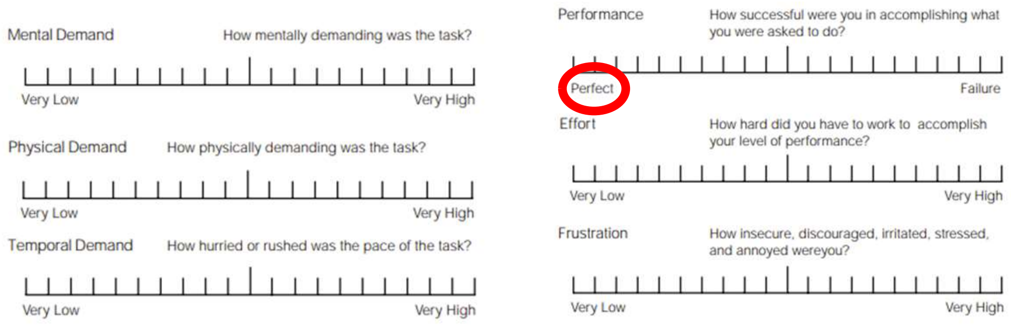
The **focus** is on measuring the "**immediate often unverbalized impressions that occur spontaneously**" (Hart and Staveland, 1988). These are difficult or impossible to observe objectively.

它的重点是测量"即时的、常常是非语言化的印象，这些印象是自发产生的"（哈特和斯塔夫兰德，1988年）。这些印象往往难以或不可能客观观察到。

NASA TLX 2任务负荷指数2

- 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.
- Since it was introduced in 1988, it has had over 8000 citations.
- It is viewed as the gold standard for measuring subjective workload.
- **Originally** it was developed as a **paper and pencil** questionnaire but there are also free apps for iOS and Android
- 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: -**Mental Demand -Physical Demand -Temporal Demand -Performance -Effort -Frustration

精神需求；身体需求；时间需求；性能；努力；沮丧



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

1.Mental demand – how much **mental and perceptual** activity was required?
2.Physical demand – how much **physical activity** was required?
3.Temporal demand(时间需求) – how much **time pressure** did the user feel due to the rate at which tasks occurred?
4.Frustration (沮丧)– how insecure, discouraged or irritated did the user feel in the task?
5.Effort – how hard did the user have to work (mentally and physically) to accomplish their level of performance?
6.Performance – **how successfully** did the user think they accomplished the task?

NASA TLX Scoring 1评分

- 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 steps** process:

1.**Identifying** the relative **importanc**e of the 6 dimensions on a user's perceived workload
2.Rating **each of the 6 dimensions** on a scale

NASA TLX Rating the dimensions

- Users mark their **score** on each of the six dimensions.
- 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. 记录右边
- The score on a dimension is calculated as the tick number (1, 21) – 1 multiplied by 5. ((1,21)-1)*5

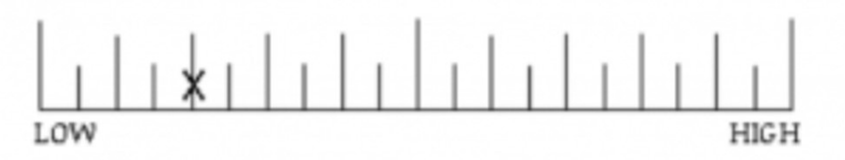
NASA TLX Relative weighting of dimensions 相对权重确定

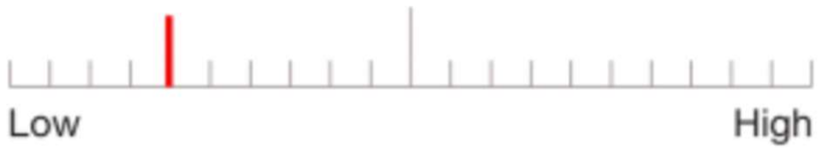
- A user reflects on the task they've been asked to perform and is shown each paired combination of the six dimensions to decide which is more related to their personal definition of workload as related to the task.
- This means a user considers 15 paired comparisons. For example, they need to decide whether Performance or Frustration "represents the more important contributor to the workload for the specific task you recently performed."
- 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.
-The sum of the weights should be 15. -----然而 下面是转折
- The relative weighting of the six dimensions is often **not** measured or used.并不测量或使用六个维度的相对权重
- Not measuring the relative weighting makes the NASA TLX **simpler to administer**.
- 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).
- When the dimensions are not rated the method is called the 'raw TLX score' 原始TLX分数（不加权）

NASA TLX Rating the dimensions 2 –计算举例

- For example, the images show the rating on a paper questionnaire (top) and on a mobile app (bottom)
- The fifth tick mark is selected, so the rating score is: (5 – 1) * 5 = 20

类比尺子





NASA TLX - What do the scores tell us?

1.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，则对每个维度的个人评分乘以它们各自的权重，然后相加并除以15，得出一个从0到100的任务的总体感知工作负荷分数。

2. 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.
如果没有使用权重，则对每个维度的个人评分相加，并除以6，得出一个从0到100的任务的总体感知工作负荷分数。

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**.

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).ASA TLX是主观工作负荷的有效衡量工具

System Usability Survey (SUS)- 系统可用性量表

- The System Usability Scale (SUS) provides a "quick and dirty", reliable tool for measuring usability. 快速简单
- It was created by John Brooke in **1986**.
- It consists of a **10 items questionnaire** with five response options for each item ranging from Strongly agree to Strongly disagree.
- It **enables the evaluation of a wide variety of products and services, including hardware, software, mobile devices, websites and applications**.各种产品和服务进行评估成为可能，包括硬件、软件、移动设备、网站和应用程序。

System Usability Survey (SUS) – benefits 好处

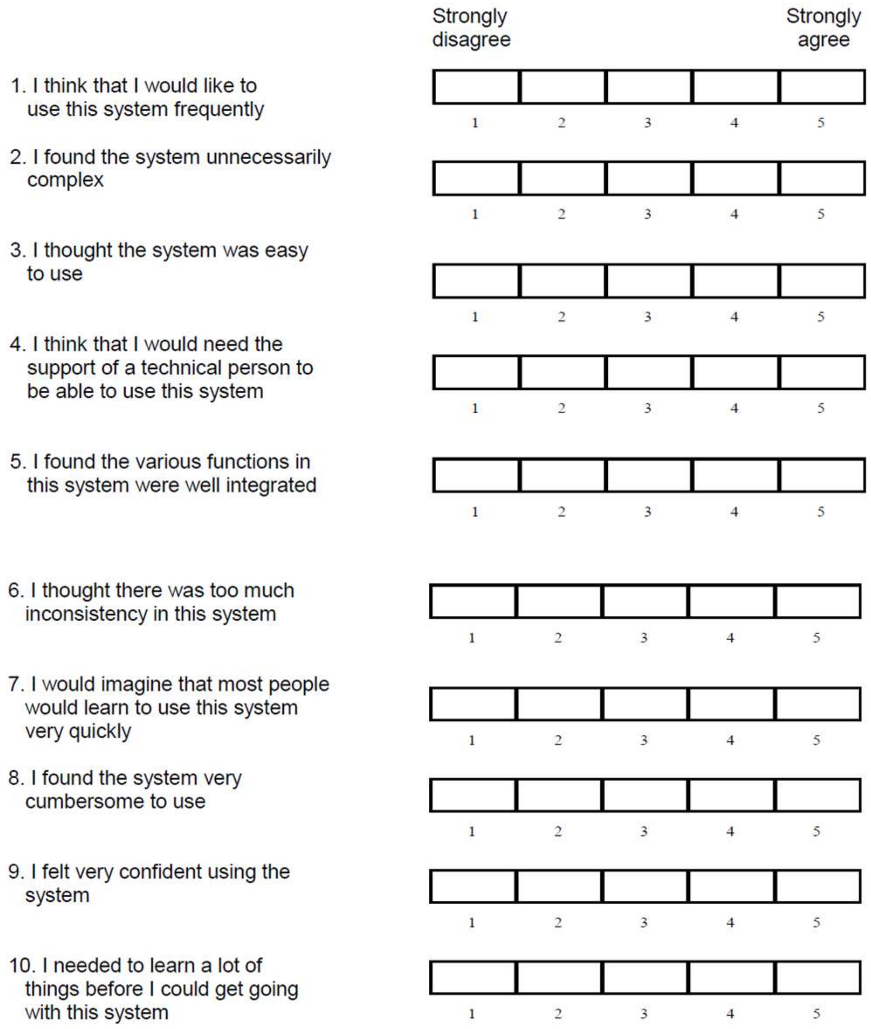
- SUS has become an industry standard行业标准, with references in over 1300 articles and publications.

The **noted benefits** of using SUS include:

- It is a **very easy scale** to administer to **participants容易评估参与者**
- It can be used on **small sample sizes** with **reliable results小样本可靠结果**
- The SUS has been validated and shown to effectively differentiate between usable and unusable systems有效区分可用&不可用系统

System Usability Survey (SUS) – scale 评分

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 a **five point Likert scale五个不一样的评分标准在十个项目里**



	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

System Usability Survey (SUS) – scoring

- The SUS is given to users when they have completed using the system which is being evaluated完成之后发调查表
- They score each of the **10 items** by marking one of the **five boxes**
- 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，代表正在研究的系统的整体可用性的综合度量。需要注意的是，**单个项目的得分本身并没有意义。**

- To calculate the SUS score, first **sum the score** contributions from each item. Each item's score contribution will range **from 0 to 4.**
- 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.
- Multiply the sum of the scores by 2.5 to obtain the overall score.
- SUS scores have a range of 0 to 100.
- Based on research, a SUS score above **68** would be considered above average and anything below 68 is below average.

Statistical tests目的： to determine if the perceived workload or system usabiity score has changed significantly

要计算SUS得分，首先要对每个项目的得分贡献进行求和。每个项目的得分贡献范围从**0到4**。
对于1、3、5、7和9号项目（奇数项目），得分贡献是**刻度位置减1**。对于2、4、6、8和10号项目（偶数项目），得分贡献是**5减去刻度位置**。
将得分总和**乘以2.5**以获取总体得分。
SUS得分的范围是**0到100**。
根据研究，SUS得分在68以上被认为是高于平均水平的，而在68以下则是低于平均水平的。**平均水平:68**

Statistical testing -统计测试

- 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.
- 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.
- To determine whether the differences in scores are significantly different we can use a statistical test为了确定得分差异是否显着不同，我们可以使用统计测试

Wilcoxon Signed Rank Test -ideal for analyzing data from Likert and other scales e.g. the NASA TLX and SUS. Wilcoxon符号秩检验，它非常适合分析来自李克特和其他量表的数据

- 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**两次**评估时使用它
- 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. 最少**5**人，但人越多结果更好识别显著差异

Wilcoxon Signed Rank Test示例

- 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/

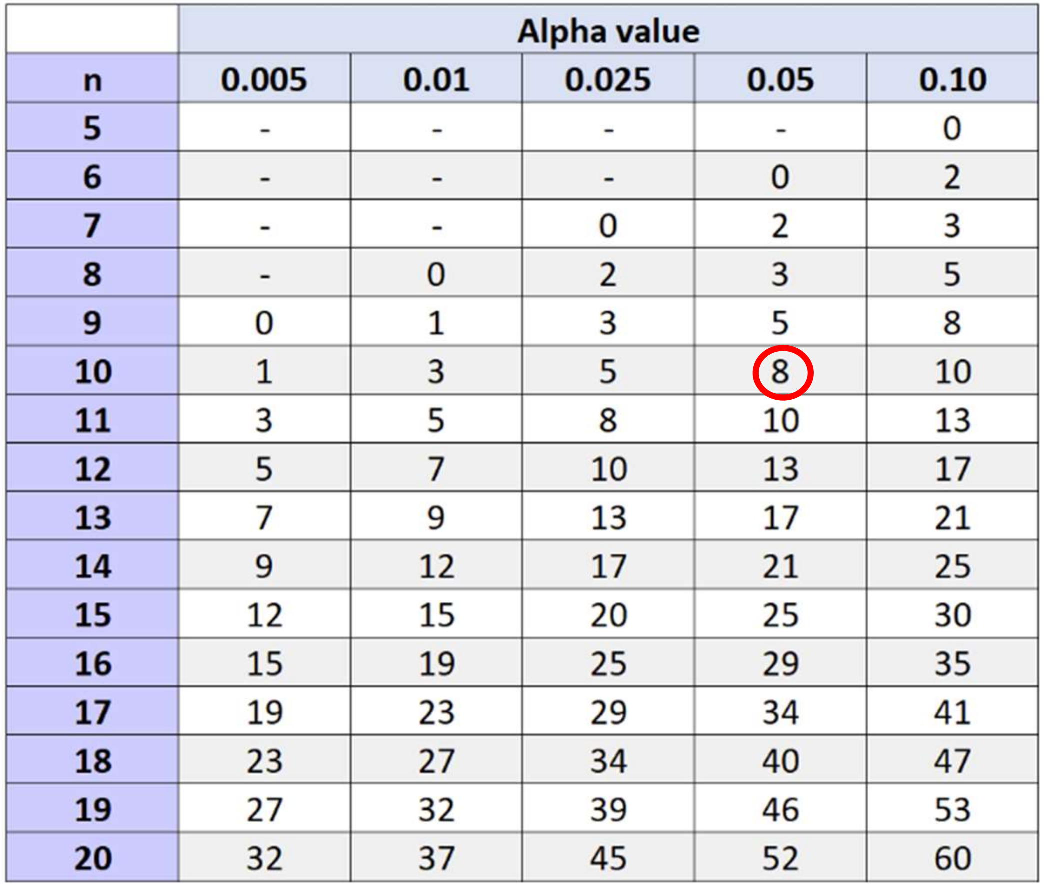
Look up the calculated **W test statistic** in the table of critical values

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

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 如果发现差异，证明95%的确定性是真的差异

Statistical testing 5

- We use an alpha value aka significance **level of 0.05**
- We find the row that corresponds to **our number of users** aka n.
- 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.



	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

Statistical testing - Mann-Whitney U test 第二个Test

- 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**
- This is known as the Mann-Whitney U test（曼-惠特尼U检验）. There is also an online calculator and you can read about the test here:

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