人机交互评估
第二部分

乔恩·伯德博士

jon.bird@bristol.ac.uk

感谢 Stuart Gray、Pete Bennett、Simon Lock、Thomas Bale、Harry Field 制作了其中一些幻灯片

图片免版税来自 www.pexels.com

今天的讲座

·问卷调查

· NASA 任务负荷指数 (NASA TLX)

·系统可用性量表（SUS）

·统计测试以确定感知是否
工作负载或系统可用性分数发生了显着变化

调查问卷 - 定义

·问卷调查要求人们
以纸质或数字方式回答问题,例如在网页或应用程序上
·它们可以大规模使用,资源需求低

·他们生成了人口统计数据的集合
数据和用户意见
·它们可用于评估设计和了解用户需求

问卷调查 - 提示

确保你提出了可行数量的问题（问题疲劳是一个问题）

·注意引导性问题,例如“为什么您在导航时遇到困难？”
·自己制作问卷很困难
·最好使用现有的调查问卷
已经过验证,即他们测量了他们声称测量的内容

·现在我向您介绍两种广泛使用的调查问卷

美国宇航局TLX

- NASA 任务负荷指数 (TLX) 是一份调查问卷，
  估计用户在使用系统时感知到的工作负载。
- 工作负载是一个复杂的结构，但本质上是指人们为使用系统而必须付出的
  精神和体力上的努力量。

- 由美国宇航局人类研究中心的桑德拉·哈特 (Sandra Hart) 开发
  表演团体和圣何塞大学的洛厄尔斯塔夫兰。

- 重点是衡量“自发发生的直接的、通常是非语言的印象”（Hart 和
  Staveland，1988）。这些是很难或不可能客观观察的。

美国宇航局 TLX 2

- NASA TLX 问卷最初是为航空用途而开发的,但后来被用于许多不同的领域,包括空中交通管制、机器人、汽车工业、医疗保健、网站设计和其他技术领域。

- 自1988年推出以来,已拥有超过8000个引文。
- 它被视为衡量主观的黄金标准工作量。

美国宇航局 TLX 3

- 最初是作为纸和铅笔开发的
  调查问卷，但也有适用于 iOS 和
  安卓
- 官方网站在这里：
  https:// humansystems.arc.nasa.gov/groups/TLX/index.php

美国宇航局 TLX 4

- NASA TLX 采用多维评级

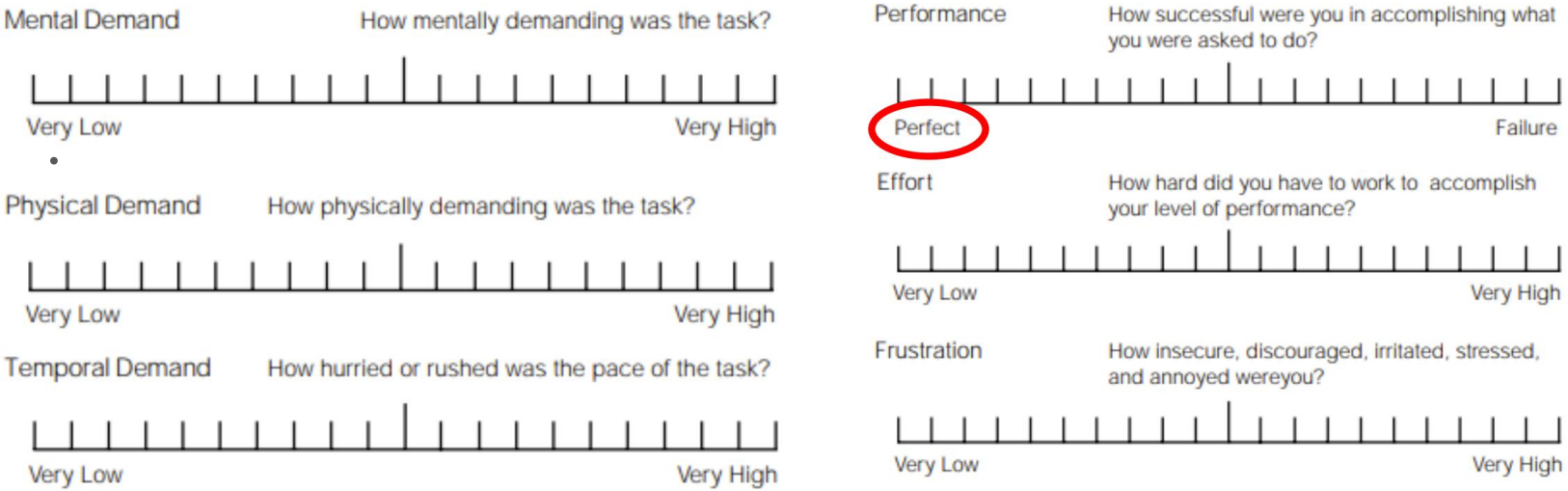
  根据六个子量表评分的加权平均值得出总体工作量分数的程序：

  - 精神需求
  - 实物需求
  - 时间需求
  - 表现
  - 努力
  - 沮丧

美国宇航局 TLX 5

- 脑力需求　需要多少脑力和知觉活动？
- 体力需求　需要多少体力活动？
- 时间需求　用户因工作而感到多少时间压力
  任务发生的速率？
- 挫败感　用户在工作中感到多么不安全、沮丧或恼怒。
  任务？
- 努力程度　用户需要付出多大的努力（精神上和体力上）才能达到
  达到他们的绩效水平？
- 性能　用户认为他们完成任务的成功程度如何
  任务？

美国宇航局 TLX 6



NASA TLX 得分 1

·用户完成任务后回答NASA TLX。这是必要的,因为要求他们在任务期间完成它通常是不可能的。然而,这可能意味着用户忘记了感知到的工作负载的详细信息。

·调查问卷分两步进行评分:

1.确定 6 个维度对用户感知工作负载的相对重要性

2.在量表上对 6 个维度中的每一个维度进行评级

NASA TLX 维度的相对权重 1

·用户反思他们被要求执行的任务,并显示六个维度的每个配对组合,以决定哪个与他们对任务相关的工作负载的个人定义更相关。 ·这意味着用户考虑15 个配对比较。例如,他们需要决定“绩效”还是“挫败感”是否“代表对您最近执行的特定任务的工作负载的更重要贡献者”。

·每次将某个维度选择为更重要时,该维度的得分为1。总分是该维度的权重,范围为 0到 5。

·权重之和应为15。

NASA TLX 维度 2 的相对权重

·六个维度的相对权重通常没有被测量或
用过的。

·不测量相对权重使得 NASA TLX 更容易
管理。

·几项研究将原始TLX 分数与加权TLX 分数进行了比较,并发现了不同的结果（一些在去除权重时显示出
更好的敏感性,另一些显示没有差异,另一些则显示较低的敏感性）。

·当尺寸未评级时,该方法称为“原始 TLX”
分数

NASA TLX 尺寸评级 1

·用户在六个方面分别打分

方面。

·每个维度由一行 21

等距刻度线,将线从 0 到 100 以 5 为增量划分。如果用户

两个刻度之间的标记,然后是值

使用右勾号。

·某个维度的得分计算如下:

刻度数 (1, 21) – 1 乘以 5。

NASA TLX 评级尺寸 2

- 例如,图像显示
  对纸质调查问卷（上)和移动应用程序（下)的评分
- 选择第五个刻度线,因此评分为:(5 – 1) * 5 = 20

NASA TLX 的分数告诉我们什么？

- 如果使用权重，则每个权重的单独评级
  将各个维度乘以各自的权重，求和并除以 15，得出任务的感知工作负载总分，范围为 0 – 100。

- 如果不使用权重，则每个权重的单独评级
  维度可以相加并除以 6，从而得出 0 到 100 之间的总体感知工作负载分数。

- 6 个维度的单独评级还可以让您了解工作负载的来源。这对于希望改进设计的开发人员很有帮助。

NASA TLX 有效性

- Hart 和 Staveland 验证了这些子量表衡量的是不同的
  工作量来源。
- 随后的独立研究也发现NASA TLX 是主观工作量的有效衡量标准（Rubio 等,2004;Xiao 等,2005）。

系统可用性调查 (SUS)

·系统可用性量表（SUS)提供了“快速而肮脏”的、可靠的
衡量可用性的工具。
·由John Brooke于1986年创建。
·它由 10 项问卷组成,每项有 5 个回答选项
项目范围从“强烈同意”到“强烈不同意”。
·它可以评估各种产品和服务，
包括硬件、软件、移动设备、网站和应用程序。

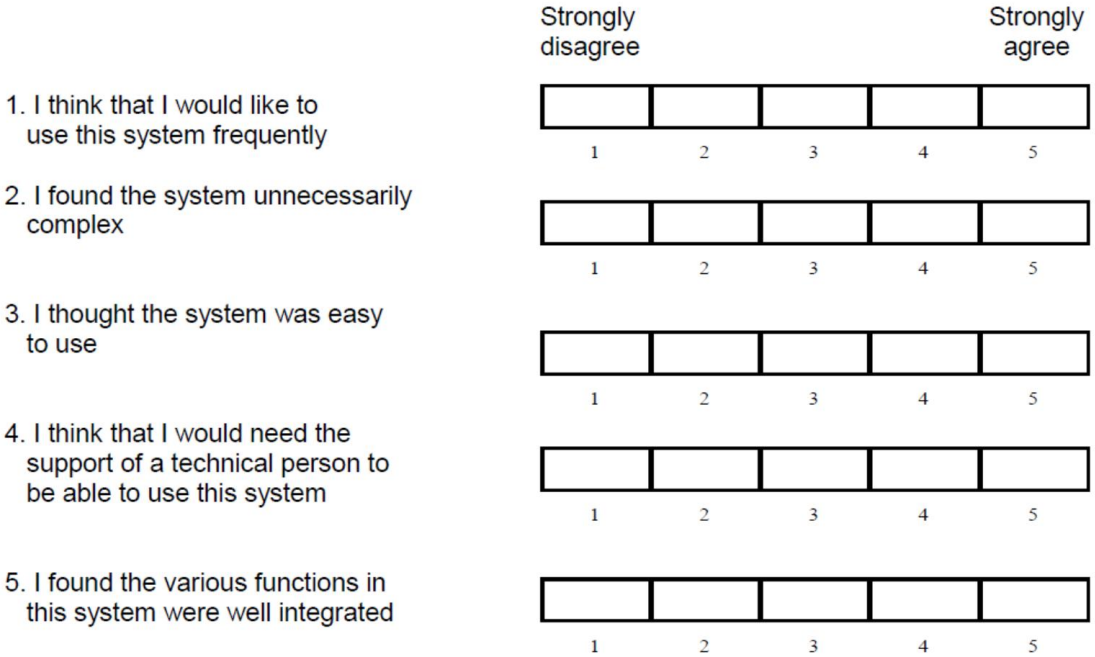
系统可用性调查 (SUS) - 优点

- SUS已成为行业标准,有参考
  发表于 1300 多篇文章和出版物。
- 使用SUS 的显着优点包括:
- 这是一个非常容易对参与者进行管理的量表
- 可用于小样本量,可靠
  结果
- SUS 已经过验证并显示可以有效地区分可用和不可用的系统

系统可用性调查 (SUS) - 规模

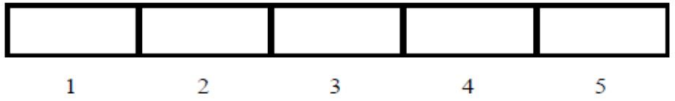
·当使用SUS时,参与者被要求对10个项目进行评分,并使用从“非常同意”到“非常不同意”的五种回答之一,即使用五点李克特量表

系统可用性调查 (SUS) – 等级 2

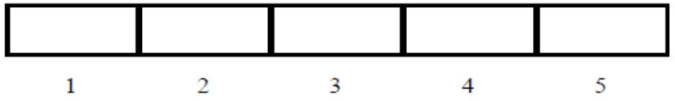


系统可用性调查 (SUS) – 等级 3

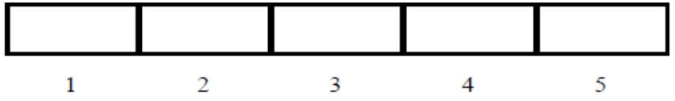
6. I thought there was too much inconsistency in this system



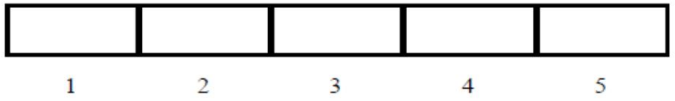
7. I would imagine that most people would learn to use this system very quickly



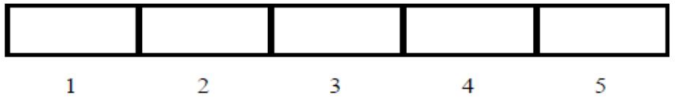
8. I found the system very cumbersome to use



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



系统可用性调查 (SUS) – 得分 1

- 当用户使用完正在评估的系统后,将向用户提供SUS

- 他们通过标记其中一项来对 10 个项目中的每一个项目进行评分
  五个盒子

- SUS 产生一个数字,代表所研究系统的整体可用性的
  综合衡量标准。请注意,单个项目的分数本身没有意义。

系统可用性调查 (SUS) – 得分 2

- 要计算SUS 分数,首先将每个项目的分数贡献相加。每个项目的分数贡献范围为 0 到 4。
- 对于第 1、3、5、7 和 9 项（奇数项)的分数贡献是标度位置减 1。对于项目 2、4、6、8 和 10 （偶数项目） ,贡献是 5 减去标度位置。

- 将分数之和乘以2.5,得到总分。
- SUS 分数的范围为0 到100。
- 根据研究，SUS 分数高于68将被视为高于平均水平,低于 68 则被视为低于平均水平。

统计检验

- 您可能会让用户对两种不同设计的SUS 进行评分,并想知道一种设计是否明显优于另一种。
- 同样,您可能想知道游戏中的两个难度级别是否显着不同,以便让用户对两个级别的工作量进行评分。

- 判断分数差异是否显着

  不同的是我们可以使用统计检验

统计检验2

- 有很多统计测试,但我将向您展示两个
  这对你的项目很有用。
- 第一个是 Wilcoxon 符号等级测试,它非常适合
  分析来自 Likert 和其他尺度（例如 NASA TLX和 SUS)的数据。

- 当一个用户进行两次评估时使用它,例如在两个不同的难度级别对游戏的工作量进行评级。

- 当用户数量较少时,这是一个很好的测试 -最少为 5 个;但是,当您拥有大量
  用户时,它可以更好地识别显着差异。

统计检验3

- 制作一个表格,其中每一行代表一个用户的分数,每个分数代表一个用户的分数。
  列一个单独的评估分数。
- 我展示了三位用户评估游戏工作量的结果
  使用 NASA TLX 有两个难度级别。
- 您至少需要 5 个,最好更多

用户身份	工作负载级别 1	工作负载级别 2
U1	25	67
U2	32	56
U3	18	43

统计检验4

·将数据输入在线计算器：

https://www.statology.org/wilcoxon-signed-rank-test-calculator/

·在临界值表中查找计算出的W检验统计量

价值观

·为此，您需要知道 N，即用户数量，以及

显着性水平，我们将其设置为 0.05

·这意味着，如果发现显着差异，则 95% 确定这是真正的差异，而不是由于

随机性

统计检验5

·我们使用 alpha 值,即显着性水平 0.05

·我们找到以下行:

对应于我们的用户数量,即 n。

·如果我们有 10 个用户,那么 W

在线计算器生成的检验统计量需要

小于8否则有

无显着差异。

	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

统计检验6

- 如果我们比较两个不同群体（例如经验丰富的游戏玩家和新手游戏玩家）生成的两组值，那么我们会使用不同的测试来查看它们是否有显着差异

- 这称为Mann-Whitney U 检验。

  还有一个在线计算器，您可以在此处阅读有关测试的信息：

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

阅读

·阅读有关 NASA TLX 的原始论文：

SG 哈特和 LE 斯塔夫兰 (1988)。

NASA-TLX（任务负荷指数)的发展:实证和理论研究的结果。进展中

心理学（第 52 卷,第 139-183 页）。

北荷兰省。

·阅读原版SUS论文

·了解有关 Wilcoxon 签名排名的更多信息测试

下周研讨会之前

·请复习讲座

NASA TLX 上的材料和

他们的

·您的研讨会活动将

涉及使用这两种技术评估您的游戏