南京大学
NANJING UNIVERSITY

人机交互的软件工程方法
—— 评估之用户测试

主讲教师：冯桂焕
fgh@software.nju.edu.cn
2012年春季

背景

- 用户测试
  - 在受控环境中（类似于实验室环境）测量典型用户执行典型任务的情况
  - 目的是获得客观的性能数据，从而评价产品或系统的可用性，如易用性、易学性等
  - 最适合对原型和能够运行的系统进行测试
  - 可对设计提供重要的反馈
  - 在可用性研究中，往往把用户测试和其他技术相结合

测试设计

- 用户测试须考虑实际限制并做出适当的折衷
  - 应确保不同参与者的测试条件相同
  - 应确保评估目标特征具有代表性
  - 实验可重复，但通常不能得到完全相同的结果
  - 以DECIDE框架为基础
- 1：定义目标和问题
  - 目标描述了开展一个测试的原因，定义了测试在整个项目中的价值
  - 目标是对关注点的说明和解答
    - 举例：对菜单结构的关注
    - 用户在第一次尝试使用时将能选择正确的菜单
    - 用户在少于5秒的时间内，能够导航到正确的3级菜单

- 2：选择参与者
  - 参与者的选择对于任何实验的成功至关重要
  - 了解用户的特性有助于选择典型用户
    - 要尽可能接近实际用户
  - 通常也需要平衡性别比例
  - 至少4~5位，5~12位用户就足够了
- 参与者安排
  - 各种实验情形的参与者不同
  - 各种情形的参与者相同
  - 参与者配对

- 参与者不同
  - 随机指派某个参与者组执行某个实验情形
  - 缺点
    - 要求有足够多的参与者
    - 实验结果可能会受到个别参与者的影响
      - 解决：随机分配or预测试
  - 优点
    - 不存在“顺序效应”
    - 即参与者在执行前一组任务时获得的经验将影响后面的测试任务

- 参与者相同
  - 相同的参与者执行所有实验情形
  - 与前一种方法相比，它只需一半的参与者
  - 优点
    - 能够消除个别差异带来的影响
    - 便于比较参与者执行不同实验情形的差异
  - 缺点
    - 可能产生“顺序效应”
    - 解决方法：均衡处理
      - 如果有两项任务A和B，那么，应让一半的参与者先执行A，再执行B，另一半则先执行B，再执行A

- 参与者配对
  - 根据用户特性（如技能和性别等），把两位参与者组成一组，再随机地安排他们执行某一种实验情形
  - 适用于参与者无法执行两个实验情形
  - 缺点
    - 实验结果可能会受一些未考虑到的重要变量的影响
    - 如在评估网站的导航性能时，参与者使用互联网的经验将影响实验结果
    - 因此，“使用互联网的经验”即可作为一个配对标准

几种安排方法的比较

参与者安排	优点	缺点
不同参与者	无顺序效应	需要许多参与者；可能受个别参与者的影响（可通过随机编组等方法解决该问题）
相同参与者	能消除各种实验情形下的个体差异	需要均衡处理以避免顺序效应
配对参与者	无顺序效应；能消除个别差异的影响	可能忽略一些重要变量，造成配对不当

- 3：设计测试任务
  - 测试任务应当与定义的目标相关
  - 测试任务通常是简单任务
    - 如查找信息
  - 有时采用较为复杂的任务
    - 如加入在线社团等
  - 任务不能仅限于所要测试的功能，应使用户全面的使用设计的各个区域
    - 如关注搜索功能的可用性，可请求参与者搜索找出产品X
    - 更好的方法就是请求参与者找出产品X并同产品Y进行比较
  - 每项任务的时间应介于5~20分钟
  - 应当以某些合乎逻辑的方法安排任务
    - 开始时，先提出简单问题有助于增强用户的自信心

- 4：明确测试步骤
  - 在测试之前，准备好测试进度表和说明，设置好各种设备
  - 正式测试前应进行小规模测试
  - 在必要时，评估人员应询问参与者遇到了什么问题
  - 若用户确实无法完成某些任务，应让他们继续下一项任务
  - 测试过程应控制在1小时之内
  - 必须分析所有搜集到的数据

- 5：数据搜集
  - 确定如何度量观测的结果
  - 使用的度量类型依赖于所选择的任务
  - 定量度量和定性度量
- 常用的定量度量
  - 完成任务的时间
  - 停止使用产品一段时间后，完成任务的时间
  - 执行每项任务时的出错次数和错误类型
  - 单位时间内的出错次数
  - 求助在线帮助或手册的次数
  - 用户犯某个特定错误的次数
  - 成功完成任务的用户数

6：数据分析

- 变量
  - 实验的目的是回答某个问题或测试某个假设，从而揭示两个或更多事件之间的关系
  - 这些“事件”称之为“变量”
- 自变量
  - 为回答假设问题，需被操作的一个或多个变量
    - 即开始实验之前，已经设置好的变量
  - 复杂的实验可能包含不止一个自变量
    - 如假设用户的反应速度不仅取决于菜单选项的数目，也取决于菜单中应用的命令选择

- 因变量
  - 能在实验中测量的变量
  - 其值依赖于自变量的变化
  - 如：完成任务所花费的时间、出错的数目、用户偏爱和用户执行的质量
- 举例：
  - 实验目标：若不用12点阵的仿宋体，而改用12点阵的楷体，那么阅读一屏文本的时间是否相同？
  - 自变量
    - 上例中的“字体”
  - 因变量
    - 上例中“阅读文本的时间”

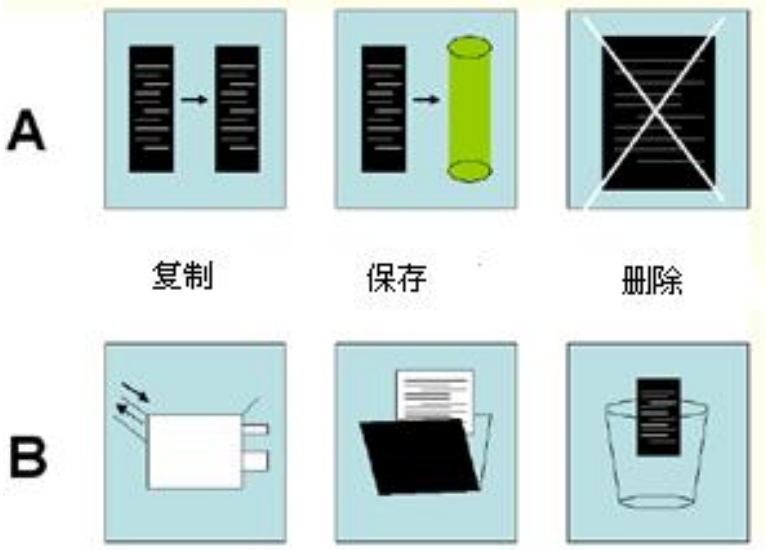
- 变量
  - 离散变量
    - 取有限数目或有限级的值
    - 如颜色可以用红、绿和蓝表示
  - 连续变量
    - 可能有上限和下限，不过可以取该范围内的任何值
    - 如：一个人的高度或完成一项任务所花费的时间
    - 能通过分段取值而离散化
      - 如：将身高分成低（<1.5m）、中（1.5~1.8m）和高（>1.8m）

测试准备

- 建造一个测试计划时间表
  - 协调参与者的日程计划、小组成员的日程计划及实验室的可使用性
- 在测试过程中编写对应的脚本
  - 脚本应当包括协调者和参与者交互的每一个方面，也应当包括一些意外事件
    - 如参与者感觉有点灰心丧气，原型出现错误等
- 安排示范性测试（Pilot test）
  - 测试可以在特定实验室里完成，也可以借助简陋的测试设备完成
  - 应当使用一个客观的参与者

图标设计评估实例-略

- 背景
  - 为一个文件处理软件包设计一个新的界面，需要用图标提供展示
  - 考虑应用两种图标设计形式
    - 自然的图像（基于纸质文档象征）
    - 抽象图像
- 目标
  - 想知道哪一个设计使用户更容易记忆
- 假设
  - 自然图标更容易记忆



- 自变量
  - 图标的形式
  - 自然的和抽象的
- 因变量
  - 关心用户记忆精确性方面的性能，还是记忆速度方面的性能，还是用户偏爱等主观度量？
  - 假设选择一个图标的速度是记忆容易程度的一个指标
    - 在选择中错误的数目
    - 选择一个图标所花费的时间

- 实验控制
  - 使观察到的任何差别清晰地归结于自变量
  - 使得对于因变量的度量是可比较的
  - 提供一个界面，除图标设计外，其他内容确定
  - 设计对每一个条件都能重复的选择任务
    - 要选择适当的图标提示
- 实验细节
  - 界面设计
  - 向用户提交一项任务（如“删除一个文件”），要求用户选择适当的图标
  - 为避免图标位置对学习的影响，在每次表示中每组图标位置的排列是随机变化的
  - 为避免学习的转移，将用户分成两组，每组采用不同的开始条件
  - 对于每个用户，测量完成任务的时间和所犯错误的数目……

参与者编号	表示标记	(1) 自然的 (s)	(2) 抽象的 (s)	(3) 参与者的平均值	(4) 自然的 (1) ~ (3)	(5) 抽象的 (2) ~ (3)
1	AN	656	702	679	-23	23
2	AN	259	339	299	-40	40
3	AN	612	658	635	-23	23
4	AN	609	645	627	-18	18
5	AN	1049	1129	1089	-40	40
6	NA	1135	1179	1157	-22	22
7	NA	542	604	573	-31	31
8	NA	495	551	523	-28	28
9	NA	905	893	899	6	-6
10	NA	715	803	759	-44	44
均值（μ）		698	750	724	-26	26
方差（σ）		265	259	262	14	14
			s.e.d. 117		s.e. 4.55	
学生的 t			0.32 (n.s.)		5.78 (p<1%，两位小数)	

网站评估实例

- 在对MEDLINEplus网站进行启发式评价后
  - 发现了可用性问题，对网站做了修改
  - 现计划对网站进行用户测试
- 1：定义目标和问题
  - 信息分类方法是否有效
  - 用户能否进退自如并且找到需要的信息
- 2：选择参与者
  - 通过问卷了解年龄、使用互联网的经验、查找医药信息的频度
    - 挑选每个月使用互联网超过两次的人员
  - 9位来自测试中心所在地的医护人员
    - 符合可用性专家所建议的5-12位
    - 7名是女性
  - 预先声明要测试NLM的一个产品

- 3：设计测试任务
  - 问题选自网站用户最经常提出的一些问题
  - 设计了5项任务
    - 任务1：查找信息，了解肩膀上的黑痣有没有可能是皮肤癌
    - 任务3：查找信息，了解是否有丙肝疫苗
    - 进行了小规模试验以确定任务的有效性
- 4：明确测试步骤
  - 准备统一的说明稿，分为五个部分
    - 以保证每一位参与者都得到相同的信息和相同的对待
  - 测试在实验室环境中进行

部分一

参与者抵达后使用

感谢你参与这项研究。

这项研究的目的是评估 MEDLINEplus 网站的界面。我们将总结评估结果，并把它提交给开发这个网站的国家医药图书馆。你使用过这个网站吗？

我们将要求你使用 MEDLINEplus 查找一些具体的医药信息。在查找信息时，请“说出”你的想法。

我们将只拍摄计算机屏幕的情况，不会拍摄你的面容。我们也将进行录音，记录你在查找过程中所说的话。我们会为你的身份保密。

一下请阅读并签署一份协议书。若有任何问题请随时提出（协议书见附表 A）。

协议书

本人在此声明：本人已年满 18 岁，并且愿意参加由 XX 及 XX 主持的 XX 研究项目。

该研究项目的目的是评估 XX 系统的可用性。XX 系统是由 XX 开发的 XX 系统，用于 XX。

测试方法是使用该系统并接受观察。本人将使用 XX 系统执行特定任务，也将回答 XX 系统以及个人使用体验相关的各种问题。

这项研究所搜集到的所有信息属于机密，任何时候都不能公开本人的身份。

本人有权利随时提出任何问题，或者随时退出测试，而不必承担任何形式的赔偿。

参与人签名　　日期（摘自文献[Cogdill 1999]）

部分二：就坐后，解释测试目的和步骤

我们先简要介绍 MEDLINEplus 网站。这是由国家医药图书馆开发的互联网产品，其目的是要帮助用户通过互联网查询权威性的医药信息。

这项研究的目的是检查 MEDLINEplus 的界面，找出有待改进的特征。同时，我们也希望了解哪些特征对用户特别有用。

几分钟之后，我们将为你安排 5 项任务。每项任务都是使用 MEDLINEplus 查找医药信息。需要指出的是，当你使用 MEDLINEplus 查找每项任务的信息时，我们的测试目标是 MEDLINEplus 的界面，而不是你本身。

你可以以正常、舒适的速度执行每项任务。我们将记录你完成每项任务的时间，但不必感到有压力，请使用正常的操作速度。如果某项任务的时间超过 20 分钟，那么请继续下一项任务。浏览器上的“主页”按钮已被设置为 MEDLINEplus 的主页。在开始执行新任务之前，请单击这个按钮，回到 MEDLINEplus 的主页。

在执行每项任务时，请设想这些信息是你或你的亲友想要了解的信息。

所有答案都可以通过 MEDLINEplus（或者它所指向的网站）找到。如果你觉得无法完成某项任务并且想中止这项任务时，请告诉我们，然后继续下一项任务。

开始之前，有什么问题吗？

部分三

执行任务前说明

在开始执行任务之前，请先用 10 分钟时间熟悉 MEDLINEplus 网站。

在熟悉网站的过程中，请说出你的想法，即，当你遇到 MEDLINEplus 的不同特征时，请告诉我们你在想什么。

你可以自由探索任何感兴趣的问题。

如果你提前完成了这个过程，请告诉我们，我们将立即进行测试任务。再次说明，当你在探索 MEDLINEplus 网站时，请告诉我们你的想法。

- 部分四
  - 若参与者忘记说出想法或不知所措时提示用

在开始使用 MEDLINEplus 查找信息之前，请读出这项任务。

完成每项任务之后，请单击“主页”按钮回到 MEDLINEplus 的主页。

提示：“你在想什么？”

“你是否不知道该怎么办？”

“请告诉我们你在想什么。”

[如果时间超过 20 分钟：“请跳过这项任务，继续下一项任务。”]

- 部分五
  - 所有任务完成后，询问参与者对某些问题的看法

你对自己执行这些任务的表现有何看法？

请说明你为什么会[遇到某个问题、出错或超时]？

你觉得 MEDLINEplus 界面的最好的方面是什么？

你觉得 MEDLINEplus 界面的最差的方面是什么？

- 5：数据搜集
  - 评估小组事先设定了成功完成每项任务的标准
  - 记录用户执行任务的全过程
    - 以下为参与者A在执行第一项任务时访问的资源

数据库
主页
MEDLINE/ 医药文献/“黑痣”
MEDLINE/ 医药文献/“痣”
主页
词典
外部网站：在线医学词典
主页
健康话题
黑素瘤
外部网站：美国癌症学会

- 数据来源
  - 根据录像和交互记录计算用户执行任务的时间
  - 问卷调查和询问阶段搜集到的数据
- 数据列表
  - 开始时间及完成时间
  - 搜索时访问的网页及数量
  - 搜索时访问的医药文献
  - 用户的搜索路径
  - 用户的负面评论和特殊的操作习惯
  - 用户满意度问卷调查数据

6：数据分析

- 网站的结构，如专栏的安排、菜单的深度和链接的组织等
- 浏览的有效性，如菜单的使用、文字密度等。
- 搜索特征，如搜索界面、提示、术语的使用是否满足一致性要求

参与者	执行时间	结束任务的原因	MEDLINEplus网页	访问外部网站	MEDLINEplus搜索	MEDLINEplus医药文献搜索
A	12	成功完成	5	2	0	2
B	12	参与者要求中止	3	2	3	0
C	14	成功完成	2	1	0	0
D	13	参与者要求中止	5	2	1	0
E	10	成功完成	5	3	1	0
F	9	参与者要求中止	3	1	0	0
G	5	成功完成	2	1	0	0
H	12	成功完成	3	1	0	6
I	6	成功完成	3	1	0	0
M	10		3	2	1	1
SD	3		1	1	1	2

- 几个问题
  - 为什么使用字母代表用户？
    - 不应透露参与者的姓名
  - “执行时间”与“结束任务的原因”有何关系？
    - 对于成功完成的任务，执行时间介于5~14分钟
    - 对于半途中止的任务，执行时间介于9~13分钟
  - 其余数据说明了什么？
    - 用户可以采取多种方式成功地完成任务
    - 如参与者A和C使用了不同在线资源

- 7：总结、报告测试结果
  - 主要问题是访问外部网站较为困难
  - 分析搜索过程
    - 有几位参与者在“健康话题”中查找不同类型的癌症时遇到了困难
  - 问卷调查结果
    - 参与者对MEDLINEplus的评价是中性的
    - 非常易学，但不易于使用
    - 在返回前一个屏幕时会遇到问题

小结

- 用户测试的适用范围
- 用户测试步骤
  - 各步骤文档的包含内容
- 能进行简单的数据分析
- 能设计和组织一个用户测试