南京大学
NANJING UNIVERSITY

# 人机交互系统

## 用户测试

主讲教师：冯桂焕

## ■评估方法二

- ●基于DECIDE框架的评估

# DECIDE评估框架

南京大学 NANJING UNIVERSITY

- 六个步骤
  - 决定评估需要完成的总体目标
  - 发掘需要回答的具体问题
  - 选择用于回答具体问题的评估范型和技术
  - 标识必须解决的实际问题，如测试用户的选择
  - 决定如何处理有关道德的问题
  - 评估解释并表示数据

# 1. 确定目标

南京大学 NANJING UNIVERSITY

- 评估目标决定了评估过程，影响评估范型的选择
- 为什么要评估?
  - 产品设计是否理解了用户需要?
  - 为概念设计选择最佳隐喻?
  - 界面是否满足一致性需要?
  - 探讨新产品应做的改进?
- 举例
  - 设计界面时，需量化评价界面质量
    - 适合进行可用性测试
  - 为儿童设计新产品时，要使产品吸引人
    - 适合采用实地研究技术，观察儿童交谈

## 2. 发掘问题

- 根据目标确定问题
  - 目标：找出为什么客户愿意通过柜台购买纸质机票，而非通过互联网购买电子机票
  - 问题
    - 用户对新票据的态度如何
      - 是否担心电子机票不能登机
    - 用户是否能够通过互联网订票
    - 是否担心交易的安全性
    - 订票系统的界面是否友好
      - 是否便于完成购票过程
- 问题可逐层分解

## 3. 选择评估范型和技术

- 范型决定了技术类型
- 必须权衡实际问题和道德问题
  - 最适合的技术可能成本过高
  - 或所需时间过长
  - 或不具备必要设备和技能
- 可结合使用多种技术
  - 不同技术有助于了解设计的不同方面
  - 不同类型数据可从不同角度看待问题
  - 组合有助于全面了解设计的情况

# 4. 明确实际问题

- ■用户
  - ●应选择恰当的用户参与评估
    - ●能代表产品的目标用户群体
    - ●可以先做测试，确定用户技能所属的用户群
  - ●任务时间多长
    - ●20分钟休息一次
  - ●可在任务执行前，安排用户熟悉系统
- ■设施及设备
  - ●如需多少台摄像机录像，具体摆放在何位置
- ■期限及预算是否允许
- ■是否需要专门技能
  - ●没有可用性专家

# 5. 处理道德问题

南京大学 NANJING UNIVERSITY

■应保护个人隐私

- ●除非获得批准，否则书面报告不应提及个人姓名，或把姓名与搜集到的数据相联系
- ●受保护的个人资料包括健康状况、雇佣情况、教育、居所和财务状况等
- ●可在评估前签署一份协议书（IRB）

■指导原则

- ●说明研究的目的及要求参与者做的工作
- ●说明保密事项，对用户&对项目
- ●测试对象是软件，而非个人

## ■指导原则-2

- ●对测试过程的特殊要求，是否边做边说等
- ●用户可自由表达对产品的意见
- ●说明是否对过程进行录像
  - ●不能拍摄用户的面部
- ●欢迎用户提问
- ●用户有随时终止测试的权利
- ●对用户话语的使用应征得同意，并选择匿名方式

# 6. 评估、解释并表示数据

南京大学 NANJING UNIVERSITY

- 搜集什么类型的数据，如何分析，如何表示
  - 通常由评估技术决定
- 可靠性
  - 给定相同时间，不同时间应用同一技术能否得到相同结果
  - 非正式访谈的可靠性较低
- 有效性
  - 能否得到想要的测量数据
- 偏见
  - 评估人员可能有选择地搜集自己认为重要的数据
- 范围
  - 研究发现是否具有普遍性
- 环境影响
  - 霍桑效应

# 小规模试验



- 对评估计划进行小范围测试
  - 以确保评估计划的可行性
  - 如检查设备及使用说明
  - 练习访谈技巧
  - 检查问卷中的问题是否明确
- 小规模试验可进行多次
  - 类似迭代设计
  - 测试——反馈——修改——再测试
  - 快速、成本低

# 可用性问题分级

南京大学 NANJING UNIVERSITY

- 评估结果总是可用性问题清单，以及改进建议
- 方法一：基于量化数据的分级
  - 如多少人遇到该问题，耗费多少时间等
- 方法二：问题严重性的主观打分，取平均值
  - 0：不是一个可用性问题
  - 1：一个表面的可用性问题
    - 如果项目时间不允许，可不予纠正
  - 2：轻微的可用性问题
    - 优先级较低
  - 3：重要可用性问题
    - 需要重视，给以高优先级
  - 4：可用性灾难
    - 产品发布之前必须纠正

## ■方法三：可用性分级的两个因素

- ●多少用户会遇到这个问题
- ●用户受该问题影响的程度

| 问题对用户的影响程度 \ 遇到问题的用户比例 | 少 | 多 |
|---|---|---|
| 小 | 低严重性 | 中严重性 |
| 大 | 中严重性 | 高严重性 |

## ■方法四：该问题只在第一次使用时出现，还是会永远出现

- ●举例：菜单条中的下拉菜单
  - ●用户从不尝试下拉用图标表示的菜单
  - ●有人告诉他们后，可马上知道如何克服该不一致性问题
  - ●因此该问题不属于永久性的可用性问题

# ■用户测试

# 背景

南京大学 NANJING UNIVERSITY

■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（视情况而言）

## ■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（定性/定量）依赖于所选择的任务

## ■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

# 分析方法

南京大学 NANJING UNIVERSITY

■定量数据

●最常用的描述性统计方法是次数统计

●举例：是否认为该技术对改进命令的访问效率有帮助？

●定量数据的次数统计、平均数统计

| 回答 | 次数 | 百分比 | |
|---|---|---|---|
| 强烈反对 | 0 | 0% | 0%反对 |
| 反对 | 0 | 0% | |
| 中立 | 3 | 30% | 30%中立 |
| 赞同 | 6 | 60% | 70%赞同 |
| 强烈赞同 | 1 | 10% | |
| 总计 | 10 | 100% | |

| 用户 | 完成任务花费的时间（分:秒） | 出错次数 |
|---|---|---|
| 1# | 1:30 | 2 |
| 2# | 3:15 | 5 |
| 3# | 4:00 | 0 |
| 4# | 2:45 | 4 |
| 5# | 3:20 | 4 |
| 平均值 | 2:58 | 3 |

## ■定性数据

- ●通常按主题分类
- ●Eg.找出获得某信息的最快途径

| 预先定义的类别 | 描述 | 用户 | 对应记录中的时间 | 备注 |
|---|---|---|---|---|
| 导航的清晰程度 | 不能找到信息 | 2# | 14:23 | 用户 2 用了 5 分钟来查找信息，最终放弃了 |
| | | 4# | 10:58 | 用户 4 在正确的页面上，但没有注意到他要找的信息 |
| | 不知道点哪里 | 5# | 11:16 | 用户 5 注意到所有标签看上去都不对，所以他不停地点击所有按钮，直到找到所需信息 |
| 文字密度 | 文字太密集了，阻碍了阅读 | 4# | 6:57 | 当用户 4 发现文字篇幅很长时放弃了继续查找 |

# 总结报告

南京大学 NANJING UNIVERSITY

■将测试的结果以书面形式反馈给产品的设计人员，以便于他们对设计进一步的分析和改进

1. 标题页
2. 测试环境描述
   - 硬件、软件版本、测试场地、测试时间
3. 执行概要
   - 简要概括测试发现（几页纸）
4. 测试描述
   - 最终版的测试计划、方法、培训和任务
5. 测试用户数据
   - 以表格形式描述用户的年龄、职业、经历
6. 结果
   - 以图表形式描述花费的时间、出错次数、问卷反馈等
   - 讨论和分析，适当引用用户言论
7. 正面反馈列表
8. 针对发现问题的建议列表，按照问题的严重等级和修复的难易程度降序排列。其中，每条建议内容包括：
   - 诊断问题出现的原因
   - 给出相应屏幕截图
   - 给出严重程度等级
   - 准确指出遇到该问题的用户数量
   - 给出相应视频记录的时间戳
   - 可能的话引用用户的原话
   - 给出改进建议
9. 附录（原始数据和表格）
   - 背景问卷、协议书、测试脚本、数据收集表格、音视频记录、手工笔记等

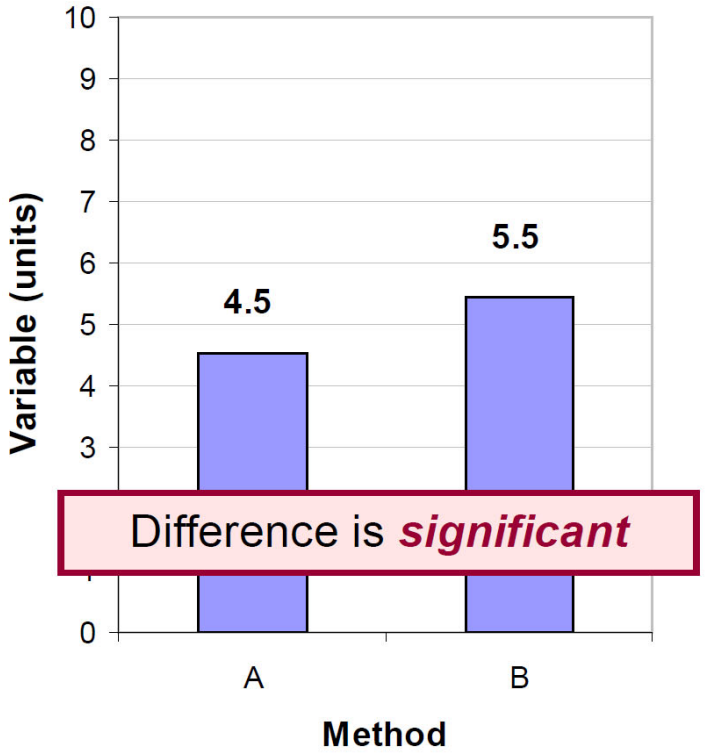
# 为什么要进行显著性检验

南京大学 NANJING UNIVERSITY

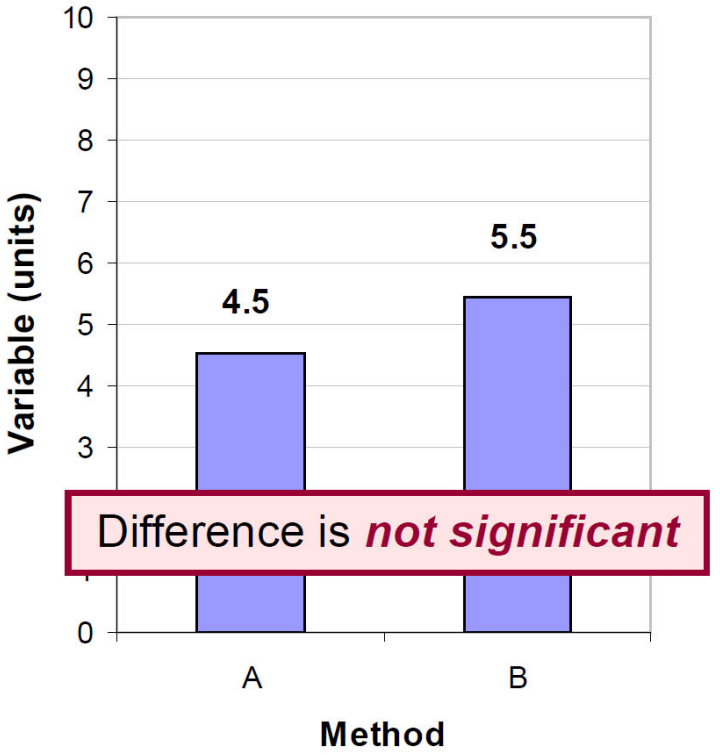
■考虑两种说法

- 1.迈克的身高是6英尺2英寸；玛丽的身高是5英尺8英寸。所以迈克比玛丽高；
- 2. 三位男性（迈克、约翰、泰德）的平均身高是5英尺5英寸；三位女性（玛丽、罗斯、杰西卡）的平均身高是5英尺10英寸。所以女性比男性高
  - 3个人的规模太小，很容易找到三位男性比女性高
  - 个性不能代表一般群体
- 当比较两个较大的群体时，没有办法收集到群体中每个个体的数据，只能“抽样”
- 显著性检验可以帮助确定我们可以将从样本群体中观察到的结果推广到整个群体的把握有多大

# 显著性检验




“显著”意味着所有可能观察到的差异是由于测试条件(方法A与方法B)导致的。




“不显著”意味着观察到的差异很可能是由偶然因素引起的。

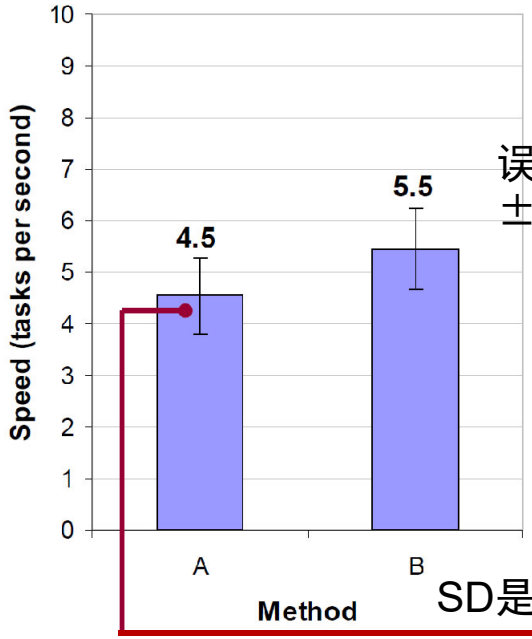
# 方差分析Analysis of Variance

南京大学 NANJING UNIVERSITY

■方差分析（ANOVA）是一种广泛应用的显著性检验方法，用来比较两组或更多组的平均值。

●当只有两个均值比较时，方差分析的计算简化为t检验。方差分析通常返回一个称为综合F的值。因此，方差分析也称为“F检验”。

第一组数据




误差条显示
±1个标准差

SD是方差的平方根

| Example #1 | | |
|---|---|---|
| Participant | Method | |
| | A | B |
| 1 | 5.3 | 5.7 |
| 2 | 3.6 | 4.8 |
| 3 | 5.2 | 5.1 |
| 4 | 3.5 | 4.5 |
| 5 | 4.6 | 6.0 |
| 6 | 4.1 | 6.8 |
| 7 | 4.0 | 6.0 |
| 8 | 4.8 | 4.6 |
| 9 | 5.2 | 5.5 |
| 10 | 5.1 | 5.6 |
| *Mean* | 4.5 | 5.5 |
| *SD* | 0.68 | 0.72 |

# 方差分析

南京大学 NANJING UNIVERSITY

■实验研究总是源于一个零假设

- ●如：方法A和方法B在打字速度上没有差别。

■实验总是先假设不同条件下的表现没有区别

- ●实验研究通常试图拒绝零假设

■请记住，通过实验研究

- ●我们收集和测试证据
- ●但不能证明任何事情

# 实验误差

南京大学 NANJING UNIVERSITY

- 第一类误差（α误差）
  - 在零假设实际上为真的情况下，却拒绝了原假设的错误
  - 统计学家将一类错误称为“轻信”的错误，会造成比现状更糟糕的情况
- 第二类误差（β误差）
  - 指的是当零假设实际上为假且应该被拒绝的情况下，却没有拒绝原假设的错误
  - 称为“无知”的错误，可能会失去改善现状的机会
- 一般我们认为，第一类误差会比第二类误差更加严重
  - 所以普遍采用一个较低的p值（0.05）来降低一类错误的风险
  - 以新型药物研发为例

# 举例：司法案例

- $H_0$：被告是无罪的
- $H_1$：被告是有罪的

| | | 陪审团裁定结果 | |
|---|---|---|---|
| | | 无罪 | 有罪 |
| 事实 | 无罪 | √ | 一类错误 |
| | 有罪 | 二类错误 | √ |

表 2.3 司法案件中的一类错误和二类错误

## ■任务完成时间的方差分析表

| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Subject | 9 | 5.080 | .564 | | | | |
| Method | 1 | 4.232 | 4.232 | 9.796 | .0121 | 9.796 | .804 |
| Method * Subject | 9 | 3.888 | .432 | | | | |

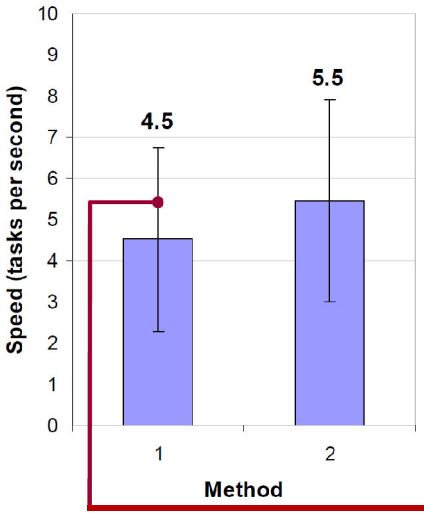
如果零假设成立，获得观测数据的概率

表述为：自变量对因变量的影响有统计学意义(F-statistic)。

Reported as…

$F_{1,9} = 9.796,\ p < .05$

Thresholds for “p”
- .05
- .01
- .005
- .001
- .0005
- .0001




| Example #2 | | |
|---|---|---|
| Participant | Method | |
| | A | B |
| 1 | 2.4 | 6.9 |
| 2 | 2.7 | 7.2 |
| 3 | 3.4 | 2.6 |
| 4 | 6.1 | 1.8 |
| 5 | 6.4 | 7.8 |
| 6 | 5.4 | 9.2 |
| 7 | 7.9 | 4.4 |
| 8 | 1.2 | 6.6 |
| 9 | 3.0 | 4.8 |
| 10 | 6.6 | 3.1 |
| *Mean* | 4.5 | 5.5 |
| *SD* | 2.23 | 2.45 |

Reported as…

$F_{1,9} = 0.626$, ns

Two ways of reporting non-significant effects:

- If $F < 1.0$, use "ns"
- If $F > 1.0$, use "$p > .05$"

| | DF | Sum of Squares | Mean Square | F-Value | P-Value | Lambda | Power |
|---|---|---|---|---|---|---|---|
| Subject | 9 | 37.372 | 4.152 | | | | |
| Method | 1 | 4.324 | 4.324 | .626 | .4491 | .626 | .107 |
| Method * Subject | 9 | 62.140 | 6.904 | | | | |

# 实验结论

南京大学 NANJING UNIVERSITY

Created using GoStats

ANOVA_table_for_Entry speed (wpm)

| Effect | df | SS | MS | F | p |
|---|---|---|---|---|---|
| Group | 1 | 73.737 | 73.737 | 0.618 | 0.4401 |
| Participant(group) | 22 | 2624.205 | 119.282 | | |
| Layout | 1 | 29664.381 | 29664.381 | 533.785 | 0.0000 |
| Layout_x_Group | 1 | 80.007 | 80.007 | 1.440 | 0.2430 |
| Layout_x_P(group) | 22 | 1222.620 | 55.574 | | |
| Trial | 4 | 1298.277 | 324.569 | 78.825 | 0.0000 |
| Trial_x_Group | 4 | 2.688 | 0.672 | 0.163 | 0.9564 |
| Trial_x_P(group) | 88 | 362.348 | 4.118 | | |
| Layout_x_Trial | 4 | 172.752 | 43.188 | 10.706 | 0.0000 |
| Layout_x_Trial_x_Group | 4 | 10.887 | 2.722 | 0.675 | 0.6113 |
| Layout_x_Trial_x_P(group) | 88 | 354.997 | 4.034 | | |

Data_file: EntrySpeed.txt

- Layout effect is significant ($F_{1,22}$ = 533.8, $p$ < .0001)
- Trial effect is significant ($F_{4,88}$ = 78.8, $p$ < .0001)
- Layout by trial interaction effect is significant ($F_{4,88}$ = 10.7, $p$ < .0001)
- Group effect is not significant ($F_{1,22}$ = 0.62, *ns*)

# 图标设计评估实例-略

■背景

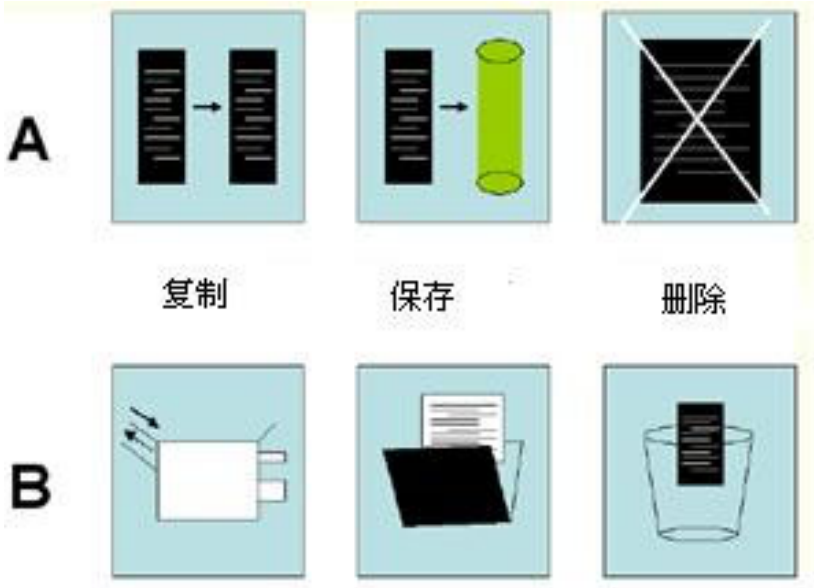
- ●为一个文件处理软件包设计一个新的界面，需要用图标提供展示
- ●考虑应用两种图标设计形式
  - ●自然的图像（基于纸质文档象征）
  - ●抽象图像

■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
- ●为避免顺序效应，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……

## ■讨论：从下表中可以得出什么结论？

| 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/ 7名是女性
    - 符合可用性专家所建议的5-12位
  - 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

## ●签署协议

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）。

# ■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

- ●任务完成之后填写调查问卷
- ●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
  - ●如必须找到并访问3-9个相关网页
- ●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



- DECIDE评估框架
  - 6个步骤
- 可用性问题分级
- 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

# 小结

南京大学 NANJING UNIVERSITY

- 常用评估范型和技术
  - 范型和技术的区别
- 技术的选择
  - 哪些影响因素
  - 为避免偏差，建议综合多个评价者的意见
    - 研究发现，一位可用性专家作出的严重性评价与真实结果之间的误差在0.5以内（5分制）的概率只有55%
    - 4名专家所做评价的平均值，其概率为95%