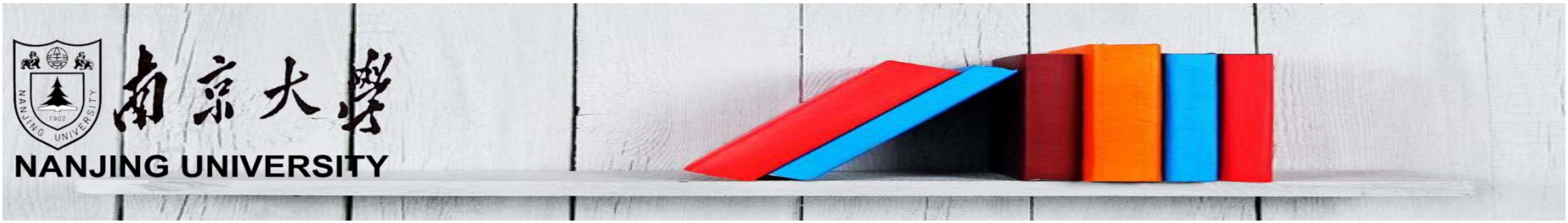


人机交互的软件工程方法
—— 评估之询问用户和专家

主讲教师：冯桂焕

2016年春季

背景

- 了解用户的需要和对产品的意见和建议
  - 观察用户
  - 询问用户
    - 适用于客观上较难度量的、与用户主观满意度和可能的忧虑心情相关的问题
    - 访谈和问卷调查
      - 在研究用户如何使用系统，以及哪些系统功能是用户非常喜欢或不喜欢的方面也非常有效
  - 不知道该怎么做或者对预期的结果没有把握
    - 请专家帮忙
      - 不能帮助大家成为可用性专家
      - 但有助于更好地去评估自己和他人的工作

询问用户之访谈

- 访谈：有目的的对话过程
  - 开放式（或非结构化）访谈，结构化访谈，半结构化访谈和集体访谈
- 指导原则
  - 避免过长的问题
  - 避免使用复合句
    - “这款手机与你先前拥有的手机相比，你觉得如何”
    - “你觉得这款手机怎么样？你是否有其他的手机？若是的话，你觉得它怎么样？”
  - 避免使用可能让用户感觉尴尬的术语或他们无法理解的语言
  - 避免使用有诱导性的问题
    - 你为什么喜欢这种交互方式？
  - 尽可能保证问题是中性的

访谈步骤

- “开始”阶段
  - 访问人先介绍自己
  - 解释访谈的原因，消除受访人对道德问题的疑虑，询问受访人是否介意被记录（录音或摄像）
- “热身”阶段
  - 先提出简单的问题
- 主要访谈阶段
  - 按逻辑次序由易到难提问
- “冷却”阶段
  - 提出若干容易的问题，消除用户的紧张感觉
- 结束访谈
  - 感谢受访者，关闭录音机，收好笔记本，表面访谈已经结束

访谈类型

- 非结构化访谈
  - 问题是开放式的，不限定内容和格式
  - 受访人自行选择详细回答还是简要回答
  - 访问人应确保能够搜集到重要问题的回答
- 结构化访谈
  - 根据预先确定的一组问题进行访谈
  - 问题通常是“封闭式”的，它要求准确的回答
- 半结构化访谈
  - 开放式问题+封闭式问题
  - 注意不要暗示答案
- 集体访谈
  - 基本思想：个别成员的看法是在应用的上下文中通过与其他用户的交流而形成的
  - “焦点小组”是集体访谈的一种形式

焦点小组

- 非正式的评估方法
  - 在界面设计之前和经过一段使用之后评估用户的需要和感受
  - 是市场、政治和社会科学研究经常使用的方法
  - 人数限制：由大约6到9个典型用户组成
  - 如在评估大学的网站时，可考虑由行政人员、教师和学生们组成3个分别的焦点小组
- 主持人工作
  - 事先列出一张讨论问题和数据收集目标的清单
  - 保持所谈论的内容不离题
  - 保证小组的每个成员都积极参与谈论
  - 讨论结果的分析报告
- 焦点小组存在风险

询问用户之问卷调查

- 问卷调查是用于搜集统计数据和用户意见的常用方法
  - 可单独使用
  - 也可与其他技术结合使用
- 问卷设计原则
  - 应确保问题明确，具体
  - 在可能时，采用封闭式问题并提供充分的答案选项
  - 对于征求用户意见的问题，应提供一个“无看法”的答案选项
  - 注意提问次序，先提出一般化问题，再提出具体问题

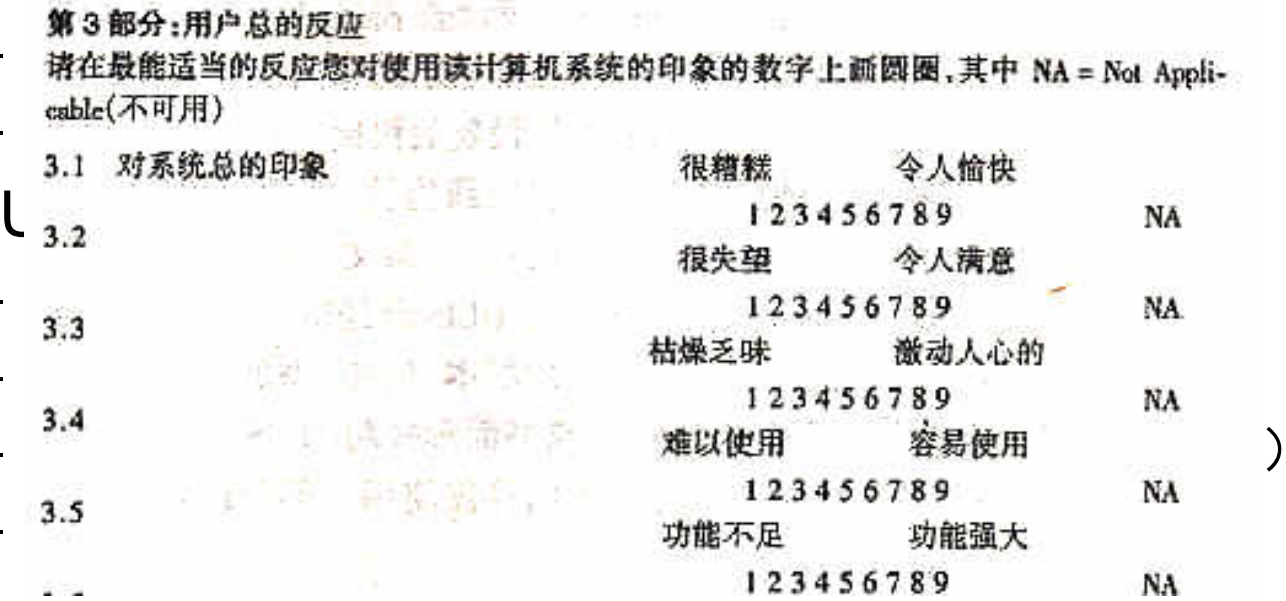
- 问卷设计原则-2
  - 避免使用复杂的多重问题
  - 在使用等级标度时，应设定适当的等级范围，并确保它们不重叠
    - 做到直观、一致
  - 避免使用术语
  - 明确说明如何完成问卷
    - 如说明应在选项前的方框内打“√”
  - 在设计问卷时，既要做到紧凑，也应适当留空

问题类型

- 常规问题
  - 年龄、性别、职业、居住地、应用计算机的经验等
- 自由回答问题
  - 如：你能够对这个界面提出改进意见吗？
  - 能够提出设计人员没有考虑到的建议
- 量化分级问题
  - 要求用户以数值尺度判断一个特定陈述
  - 如：系统容易从错误状态恢复
    - 不同意　1　2　3　4　5　同意
  - 第三章中的“Likert尺度”和“语义差异度尺度”
  - 奇数刻度较偶数刻度更常用
- 多选题
  - 对于收集用户以前的经验信息很有用

用户满意度调查表（QUIS）

- 由Ben Shneiderman开发
  - QUIS: questionnaire for user interaction satisfaction
  - –
  - –
- QU
  - –
  - –
  - –
  - –



第3部分：用户总的反应

请在最能适当的反应您对使用该计算机系统的印象的数字上画圆圈，其中 NA = Not Applicable(不可用)

3.1	对系统总的印象	很糟糕 1 2 3 4 5 6 7 8 9 令人愉快	NA
3.2		很失望 1 2 3 4 5 6 7 8 9 令人满意	NA
3.3		枯燥乏味 1 2 3 4 5 6 7 8 9 激动人心的	NA
3.4		难以使用 1 2 3 4 5 6 7 8 9 容易使用	NA
3.5		功能不足 1 2 3 4 5 6 7 8 9 功能强大	NA

)

问卷设计举例

- 比较两个不同学习系统的用户的执行情况和偏爱
  - 一个应用超媒体
  - 另一个应用顺序课程

部分 1： 对每一个系统重复

指出你同意或不同意下列陈述。(1 表示完全不同意，5 表示完全同意。)

系统在每一点告诉我做什么

不同意 1 2 3 4 5 同意

易于从故障中恢复

不同意 1 2 3 4 5 同意

当需要时很容易得到帮助

不同意 1 2 3 4 5 同意

我始终知道系统在做什么

不同意 1 2 3 4 5 同意

我始终知道我位于训练材料的什么位置

不同意 1 2 3 4 5 同意

我已经熟悉应用系统的材料

不同意 1 2 3 4 5 同意

我已经获悉有效应用一本书的材料

不同意 1 2 3 4 5 同意

我总是知道我做得怎么样

不同意 1 2 3 4 5 同意

部分 2： 比较两个系统

哪个系统（选择一个）是：

在应用中是有帮助的 A B

在应用中是有效的 A B

在应用中是我喜欢的 A B

请给你选择的一个系统添加说明。

问卷组织

- 问卷调查中的两个关键问题
  - 如何寻找有代表性的用户
  - 如何达到合理的回复率
- 有助于提高回复率的措施
  - 精心设计问卷，避免用户因为厌烦而拒绝回复
  - 参照QUIS，提供简要描述，说明用户若没有时间完成整份问卷，可以只完成简短的部分
  - 提供一个带有回复地址并粘好了邮票的信封
  - 解释为什么要进行这些问卷调查，并说明将为参与者保密
  - 在发出问卷之后，通过后续邮件、电话或电子邮件联系参与者
  - 采取一些激励措施（如有偿调查等）
  - 进行小规模测验

在线问卷调查

- 能有效而方便地搜集大量人员的意见
  - 能够快速搜集调查结果
  - 与纸张式的问卷调查相比，成本更低，甚至为零
  - 数据可以立即输入数据库进行分析
  - 可缩短数据分析的时间
  - 容易更正问卷中存在的问题
  - 回复率可能低于纸质问卷
- 两种形式
  - 基于电子邮件
    - 能够针对特定的用户，但邮件能够容纳的内容有限
  - 基于网页的调查
    - 形式灵活，并能验证数据的有效性，但调查对象是随机的

问卷调查与访谈

- 问卷调查或访谈都属于间接方法
  - 因为两者都不对用户界面本身进行研究，而只是研究用户对界面的看法
  - 都不能完全听信和采纳用户的说法
    - 询问ZAP命令的说明
    - 系统新增功能的问卷
    - 移动电话说明书的问卷
- 访谈
  - 形式更自由
  - 难以获得确切数据
  - 需要花费更多时间
  - 可在访谈后立即得到结果
  - 可能回避某些“敏感问题”的真实想法

询问专家之认知走查

- 评估应该贯穿于整个设计过程中
  - 理想情况下，系统所有实现工作开始之前就应该评估
  - 专家分析可应用于项目设计的任何阶段
- 认知走查
  - 逐步检查使用系统执行任务的过程，从中找出可用性问题
  - 无需用户参与
  - 认知走查的主要目标是确定使一个系统如何易于学习
  - 试图想象出人们在第一次使用某个产品时的想法以及所采取的动作，它的大作流程是怎样的
  - 评估的具体过程就是把用户在完成这个功能时所做的所有动作讲述成一个令人可以信服的故事

走查的步骤

- 标识并记录典型用户的特性
  - 有关用户自身心理、心理特点以及他们的知识和经验的描述
- 基于评估重点，设计样本任务
  - 应该是大多数用户要做的典型任务
- 制作界面原型（或界面描述），明确用户执行任务的具体步骤
- 由设计人员和专家级评估人员（一位或多位）共同进行分析
- 评估人员结合应用的上下文，逐步检查每项任务的操作步骤
  - 见下页
- 在完成逐步检查之后，汇总关键信息
- 修改设计，更正发现的问题

- 检查每项任务的操作步骤时，了解以下问题
  - 正确的操作对于用户是否足够明显？（可预见）
    - 即用户能否知道如何完成任务
    - Excel中换行的例子
  - 用户能否注意到正确的操作？ （可理解）
    - 功能名称或图标设计是否容易理解
  - 能否正确解释操作的响应？（可解释）
    - 执行—评估交互周期的完成
    - 网页提交按钮的例子
- 认知走查的记录工作非常重要！

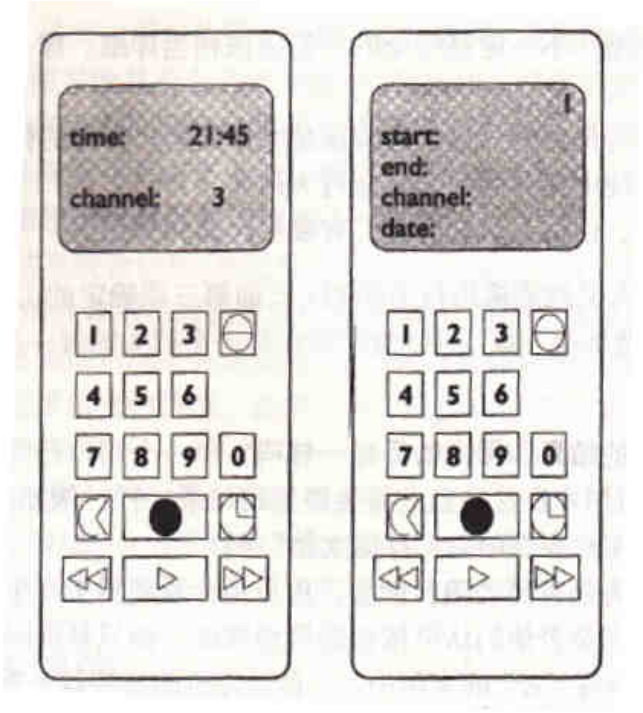
分析

- 优点
  - 不需要用户参与
  - 不需要可运行的原型
  - 能找出非常具体的用户问题
- 缺点
  - 工作量大，非常费时
  - 关注面有限
    - 只适合于评估一个产品的易学习性
    - 不太容易发现使用效率方面的可用性问题

认知走查实例

- 录像机遥控器的定时功能
  - 规划一段定时录像，录像过程在通道4上，从2005年2月24日的18:00开始，到19:15时结束

用户行为1：按“定时录像”（timed record）键
系统响应：显示转换到时间模式。在“开始：”（start:）以后出现闪烁的光标
用户行为2：按下数字1800
系统响应：显示输入的每一个数字，并且闪烁的光标移后一位
用户行为：按“定时录像”键
系统响应：闪烁的光标移动到“末端：”（end:）
用户行为4：输入数字1915
系统响应4：显示输入的每一个数字，并且闪烁的光标移后一位
……



- 对每一个行为，回答如下问题
- 用户行为1：按“定时录像”键
  - 行为的结果和用户的目标一样吗？
  - 用户能看到这个行为是可用的吗？
    - 并不清楚哪一个是“定时录像”键
  - 一旦用户找到了一个正确的行为，他们能知道这个行为是所要的吗？
  - 用户能够理解所获得的反馈吗？

协作走查

- 由用户、开发人员和可用性专家合作，逐步检查任务场景，讨论与对话元素相关的可用性问题
  - 在评估过程中，每一位专家都承担用户的角色
- 优点
  - 专注于用户任务；能够产生定量数据
  - 符合参与式设计原则
- 缺点
  - 需要各方面的专家，速度慢
  - 由于时间限制，通常只能评估有限的场景

询问专家之启发式评估

- 一种灵活而又相当廉价的评估方式
- 复习：Nielsen的十条启发式规则
  - 系统状态的可视性
  - 系统应与真实世界相符合
  - 用户的控制权及自主权
  - 一致性和标准化
  - 帮助用户识别、诊断和修复错误
  - 预防错误
  - 依赖识别而非记忆
  - 使用的灵活性及有效性
  - 最小化设计
  - 帮助及文档

启发式评估

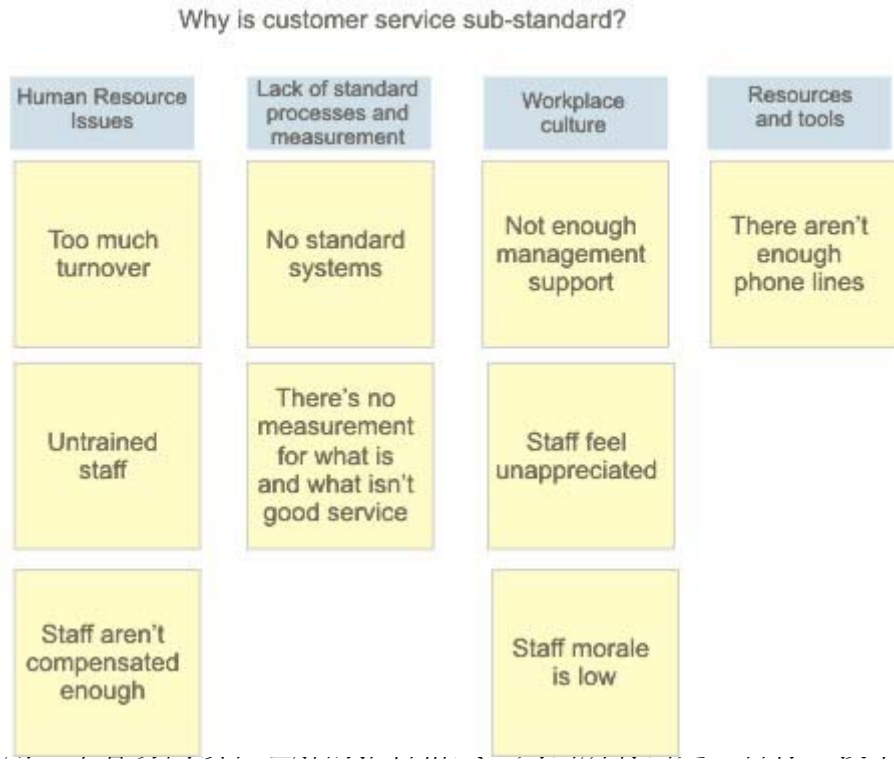
- 由可用性专家完成
- 步骤
  - 彻底检查界面
  - 将界面与启发式规则进行对比
  - 列举可用性问题
- 应用启发式规则对每一个问题进行解释与确认

问题的严重性分类

- 不同作用因素
  - 频率：有多经常？
  - 影响：有多难克服？
  - 持续时间：要多长时间克服？
- 严重性等级
  - 表面问题：不需要被修复
  - 次要问题：需要修复，但优先级较低
  - 主要问题：需要修复且优先级很高
  - 灾难性问题：必须被修复

评估步骤

阶段	步骤
准备（项目指导）	a. Figure 5 b. c. d. 、系统规格说明、用户 e. 派一个共同的记录员还
评估（评估者活动）	a. b. 所需的操作进行实际操 c. 问题，包括可能重复之
结果分析（组内活动）	a. 理解； b. 相似的问题分组； c. d. e.
报告汇总	a. 点的解释和修改建议； b. 、过程和发现。评估者可根据评估原则（heuristics）来组织发现的问题。一定要记录系统或界面的正面特性； c. 确保报告包括了向项目组指导反馈的机制，以了解开发团队是如何使用这些信息的； d. 让项目组的另一个成员审查报告，并由项目领导审定。



如何正确评估

- 分析每个问题对应的启发式规则
  - 如“主页上有太多选项”对应“审美与最小化设计”
    - 不能简单地说“我不喜欢它的颜色”
- 列出所有问题
  - 即便可能某个界面元素存在多个问题
- 至少遍历两次界面
  - 一次获得系统的初始体验
  - 另一次关注特定界面元素
- 不要局限于10条启发式规则
  - 还有各种affordances、constraints、颜色原理等

iTunes评估实例

- 评估使用的启发式规则
- 三位专家
  - 两位对Windows熟悉
  - 一位对Mac熟悉
- 问题修复的难易程度

编号	启发式规则
1	审美和最小化设计
2	有效地菜单/命令结构
3	使用简单的自然语言
4	减轻用户的记忆负担
5	一致性
6	提供反馈
7	提供清晰的出口
8	积极应对错误
9	帮助功能

等级	定义及描述
0	问题非常容易修复。在下一次版本发布之前可以由一个项目组成员完成
1	问题容易修复。涉及到特定界面元素，有明确解决方案
2	问题修复有些困难。涉及界面的很多方面，需要整个项目组成员来完成或者解决方案尚不明确
3	问题难以修复。涉及到界面的很多方面，在下一版本发布之前解决有一定难度，尚未获得明确的解决方案或是解决方案仍存有争议

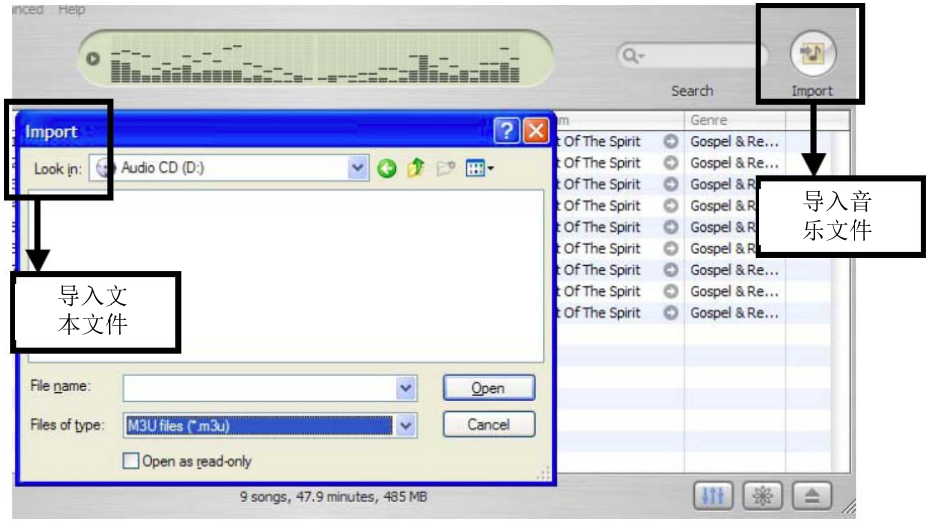
发现的问题

编号	问题描述	严重等级	修复等级	违反规则	改进建议
1	菜单和按钮不一致	3	1	5	对应用中的词汇进行分析，特别是将按钮和工具提示上的术语和菜单中具有相同功能的术语进行比较。
2	部分语言和用户的常用术语不符	3	1	3	对界面词汇进行全面分析，最终术语的确定要让真实用户参与以
3	部分按钮看上去不像是按钮	3	2	1，4，5	改变按钮颜色；当鼠标指向按钮时，高亮显示和改变颜色也会非常有用
4	部分按钮缺少工具提示	2	1	4	为所有按钮增加描述性的工具提示
5	存在一些和Windows操作规范不一致的地方	2	3	4，5	对Windows版本使用平台一致的层次化结构展现方式
6	几乎不支持撤销操作	2	3	7	当标签改变时即激活UNDO功能，对未修改内容禁止UNDO操作；同时应该支持对播放列表的UNDO操作
7	模式界面导致的不一致问题	2	3	5	该问题修改起来比较困难，可能会涉及到对界面整体布局的调整。最初可以考虑在相应按钮处提供一些下拉菜单，当某项功能不可用时就将相应菜单以灰色显示。

问题一

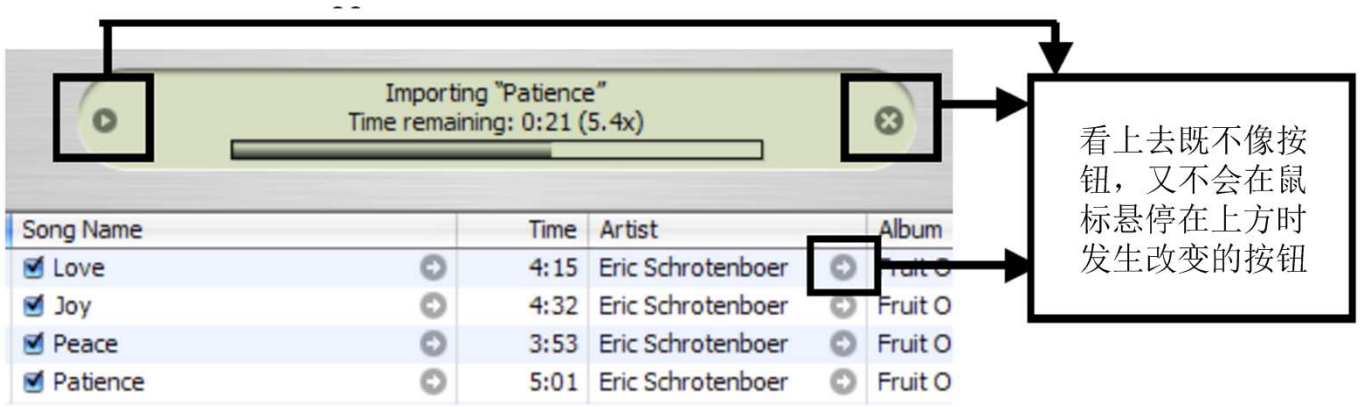
- 菜单、按钮、以及工具提示上的语言存在明显的不一致性
  - 违反第5条一致性原则

按钮/工具提示的文字	菜单文字
Import	Add File to Library
Burn Disc	Burn Playlist to Disc
Visual Effects	Visualizer



问题三

- 很多可点击的按钮看上去不像是按钮
  - 违反了第1条和第4条启发式规则：图标的形状和颜色设计应能传递图标本身的功能和作用



分析

- 启发式评估的优点
  - 不涉及用户，所以面临的实际限制和道德问题较少
  - 成本相对较低，不需要特殊设备，而且较为快捷
  - 又被称为“经济评估法”
- 启发式评估的缺点
  - 评估人员需要经过长时间的训练才能成为专家
    - 理想的专家应同时具备交互设计和产品应用域的知识
  - 可能出现“虚假警报”
    - “专家每找到一个真实的可用性问题，将发出约一个假警报（1.2），忽略大约半个问题（0.6）”

启发式评估不是用户测试

- 评估专家也不是用户
  - 尽管可能比“你”更接近典型用户
- 关系可类比
  - 代码评审vs.测试
- 启发式评估会发现用户测试容易遗漏的问题
  - 不一致的字体
  - Fitts定律问题
  - 但用户测试才是可用性的Gold Standard

友情提醒

- 邀请多个评估专家
  - 不同评估专家可能发现不同问题
  - 越多越好，但回报可能会越来越小
  - Niesen推荐3-5名评估专家
- 使用用户测试替代启发式评估
  - 不同方法发现的问题不同
  - 启发式评估更廉价
- 观察人员可以帮助评估专家
  - 只要专家已经注意到了某个问题
  - 但对用户测试而言并不适合那么做

小结

- 访谈
- 问卷调查
  - 和访谈的区别
  - 选用
- 认知走查
  - 哪些特点
- 启发式评估
  - 灵活运用

作业

- 完成对所选实践项目前期工作的系统性评估
  - 提交评估报告
  - 要求包括：评估设计、评估过程、评估结论等，要求包含用户测试
  - 分组完成

- 提交日期
  - 4月30日