



南京大學
NANJING UNIVERSITY



人机交互系统 评估的基础知识

主讲教师：冯桂焕





背景

■ 评估是设计过程的组成部分

■ 什么是评估？

- 系统化的数据搜集过程

- 用户在与原型、应用程序、计算机系统、系统组件、应用程序或界面草图交互时收集关于用户体验方面的信息，从而改进其设计

- 例如，用户能否找到特定的菜单项？图像是否有用，是否吸引人？产品是否引人入胜？

- 评估侧重系统的可用性和用户体验

■ 注意：

- 邀请用户进行评估的目的不是设法理解用户，而是评估特定用户在一个特定的环境背景中如何使用一个系统来执行一个特定的任务。



关于评估的四个“W”

■ Why-为什么要评估？

- 用户不仅期望一个可用的系统，还在追求愉悦感与参与感
 - 用户体验首先要满足客户的确切需求，其次是让用户舒心和愉快
- 对企业而言，好的设计有很大卖点；设计师可以专注于真实问题和不同用户群体的需求，而非与人们对喜爱与厌恶之处进行辩论

■ What-评估什么？

- 对象：原型、可运行系统、特定屏幕功能、完整工作流程、审美设计、安全性等等
 - 游戏开发者想要知道如何吸引用户并激发他们的兴趣，以及让用户玩游戏的时间更长
 - 玩具制造商可能会询问6岁的孩子是否可以操控玩具，是否喜欢绒毛玩具的外观，以及玩具是否安全等



问题：

■可以从哪些方面对一款个人音乐播放器（如智能手机上的APP）进行评估？

■参考答案

- 人们如何从成千上万首歌曲中选择特定曲目
- 人们能够根据自己的喜好设置播放列表
- 播放器的外观和质感如何
- 人们是否可以轻松地添加或存储新的音乐
-



关于评估的四个“W”

■ Where-在哪里评估？

- 取决于正在评估的对象
- 实验室环境：提供了必要条件系统地检查产品是否满足用户的所有要求，如智能手机按键的大小和布局的评估
- 实地环境：如小孩子是否喜欢玩新玩具以及他们玩多久会感到无聊

■ When-何时开展评估？

- 取决于产品的类型
 - 如果研发新产品，则会投入大量时间做市场调研和建立需求，之后会设计草图/故事图板，可以通过评估检查是否正确理解了用户的需求，这被称作“形成性评估”，目的是调整和完善设计
 - 评估已完成产品成功与否被称作“总结性评估”，目的是确定产品需要改进的方面



评估原则

■ 评估应该依赖于产品的用户

- 与专业技术人员的水平和技术无关

■ 评估与设计应结合进行

- 仅靠用户最后对产品的一两次评估，是不能全面反映出软件可用性的

■ 评估应在用户的实际工作任务和操作环境下进行

- 根据用户完成任务的结果，进行客观的分析和评估

■ 要选择有广泛代表性的用户

- 参加测试的人必须具有代表性





评估范型

- 快速评估
- 可用性测试
- 实地研究
- 预测性评估



快速评估

- 设计人员非正式地向用户或顾问了解反馈信息，以证实设计构思是否符合用户需要
 - 可在任何阶段进行
 - 强调“快速了解”，而非仔细记录研究发现
 - 如在设计初期了解用户对新产品的意见、在设计末期了解用户对图标设计的看法等
 - 得到的数据通常是非正式、叙述性的
 - 可以口语、书面笔记、草图、场景的形式反馈到设计过程
 - 是设计网站时常用的方法
- 基本特征：快速





可用性测试

- 20世纪80年代的主导方法
- 评测典型用户执行典型任务时的情况
 - 包括用户出错次数、完成任务的时间等
- 基本特征
 - 是在评估人员的密切控制之下实行的
- 主要任务
 - 量化表示用户的执行情况
- 缺点
 - 测试用户的数量通常较少
 - 不适合进行细致的统计分析



实地研究

- 基本特征：在自然工作环境中进行
- 目的：理解用户的实际工作情形以及技术对他们的影响

- 作用
 - 探索新技术的应用契机
 - 确定产品的需求
 - 促进技术的引入
 - 评估技术的应用

- 重难点
 - 如何不对受试者造成影响
 - 控制权在用户，很难预测即将发生和出现的情况



预测性评估

■ 研究人员通过想象或对界面的使用过程进行建模

- 专家们根据对自己对典型用户的了解预测可用性问题的可用性评估
- 逐步通过场景或基于问题回答的走查法
- 用于比较相同应用不同界面的原型法，如使用Fitts定律预测使用设备定位目标的时间

■ 基本特征

- 用户可以不在场
- 整个过程快速、成本较低





区分评估技术的因素

■评估在周期中的位置

- 设计早期阶段的评估更快速、便宜

■评估的形式

- 实验室环境or工作环境

■技术的主客观程度

- 技术越主观，受评估人员知识的影响越大，如认知走查等

■测量的类型

- 主观技术：定性数据
- 客观技术：定量数据



■提供的信息

- 底层信息：这个图标是可理解的吗？
- 高层信息：这个系统是可用的吗？

■响应的及时性

- 边做边说法可及时记录用户行为
- 任务后的走查取决于对事件的回忆

■干扰程度

- 直接观察可能会影响用户表现

■所需资源

- 设备、时间、资金、参与者、评估人员的专业技术及环境等



评估方法组合

■ 评估方法的组合取决于项目待评估的具体特性

■ 常用组合

- 启发式评估+边做边说等用户测试技术

- 专家可通过启发性评估排除显而易见的可用性问题

- 重新设计后，经用户测试，反复检查设计的效果

- 访谈+问卷调查

- 先对小部分用户进行访谈，确定问卷中的具体问题

■ 启发式评估vs.用户测试

- 前者不需要用户参与

- 二者发现的可用性问题不同，可以互补



■评估方法一

- 人机交互的实证研究方法



研究假设

- 实验通常从研究假设开始

- 假设是一种可以通过实证研究直接检验的精确问题陈述
- 一个具体的研究假设奠定了一个实验以及统计学显著性检验的基础

- 零假设 Null Hypothesis 和备择假设 Alternative Hypothesis

- 零假设通常指不同的实验条件不会产生差异；而备择假设往往是一个与零假设相反的陈述
- **实验的目标**是找到统计学证据来反驳或否定零假设，以支持备择假设



- 举例

- 例如，一个网站的开发人员正尝试弄清楚在网站的主页上是用下拉菜单还是用弹出菜单。
- H0：下拉菜单和弹出菜单在定位页面的时间开销上没有差异；
- H1：下拉菜单和弹出菜单在定位页面的时间开销上存在差异。
(备择假设和零假设应该是互斥的)
- 可以同时研究多对零假设和备择假设
 - H0：下拉菜单和弹出菜单在用户满意度评价上没有差异；
 - H1：下拉菜单和弹出菜单在用户满意度评价上存在差异。



研究假设

- 一个成功的实验，从一个或多个**好的假设**开始是至关重要的
 - 用精确而清晰的语言提出；
 - 专注于一个可以在单次实验中检验的问题；
 - 明确说明实验的对照组或实验条件。



因变量和自变量

- 一个定义明确的假设会明确说明研究的因变量和自变量。

➤ 自变量**Independent Variables**是指研究者感兴趣的的因素或因变量变化的可能“原因”

- “**Independent**” 用于说明该变量与受试者的行为无关，即参与者无法对自变量施加任何影响
- 一个自变量至少需要2个不同的取值，这些取值被称为实验条件**test conditions**
- 人的特性是天然的自变量，如年龄、性别、身高等，但它们不能被“操纵”

Factor (IV)	Levels (test conditions)
Device	mouse, trackball, joystick
Feedback mode	audio, tactile, none
Task	pointing, dragging
Visualization	2D, 3D, animated
Search interface	Google, custom



因变量和自变量

- 因变量Dependent Variables是指研究者感兴趣的结果或效果，其中术语“因”用于说明该变量依赖于受试者的行为或自变量的变化。
 - 如：任务完成时间、速度、准确性、错误率、任务重试的次数、按退格键次数等。
 - 因变量必须被明确定义！
- 研究者希望找出自变量的变化是否会引起因变量的变化，以及如何引起因变量的变化。
 - 研究必须是可复现的！



因变量和自变量

• 如何区分

- 自变量通常是研究人员可以控制的实验条件；因变量通常是研究者需要测量的实验结果
- 举例：下拉菜单和弹出菜单在定位页面的时间开销上没有差异。找出自变量和因变量。

• 典型技术相关自变量

- 不同类型的技术或设备。如打字与语音听写，鼠标与操纵杆、触摸板等指向设备等。
- 不同类型的设计。如下拉式菜单与弹出式菜单，字体大小，对比度，背景色，网站架构等。

● 典型因变量

- 效率、准确性、主观满意度、易学性和易记性、体力或认知需求。



实验构成

- 实验条件：指的是我们需要比较的不同技术、设备或程序；即变量取值
- 实验单位：指的是我们应用实验条件的对象，在人机交互研究领域，实验单位通常是具有特定特征的人类受试者，特定特征例如性别、年龄、计算机使用经验等；
- 分配方式：指的是将实验单位分配到不同实验条件的方式。
- 举例：
 - 比较传统**QWERTY**键盘与**DVORAK**键盘的打字速度
 - 实验条件就是键盘的类型：**QWERTY**键盘或**DVORAK**键盘
 - 为了达到公平比较的目的，研究人员需要要求受试者之前都没有使用这两种键盘的经验
 - 分配—抛硬币：如果是正面，受试者就被分配到**QWERTY**条件下；否则，受试者就被分配到**DVORAK**条件下



分配方式

- 实验性研究的强大在于它能够探索因果关系，能够实现这一目标的主要原因即在于它的**完全随机化**
 - 随机化即是将实验条件随机分配给实验单位
- 传统的随机化方法包括：掷硬币、掷骰子、轮盘赌、从盒子里取球等，目前已较少采用
- 另一种方法是使用随机数字表
- 在一个设计良好的实验中，你经常会发现，你不仅需要随机分配实验条件，还需要随机分配其他因素

行号	随机数				
000	10097	32533	76520	13586	34673
001	37542	04805	64894	74296	24805
002	08422	68953	19645	09303	23209
003	99019	02529	09376	70715	38311
004	12807	99970	80157	36147	64031
005	66065	74717	34072	76850	36697

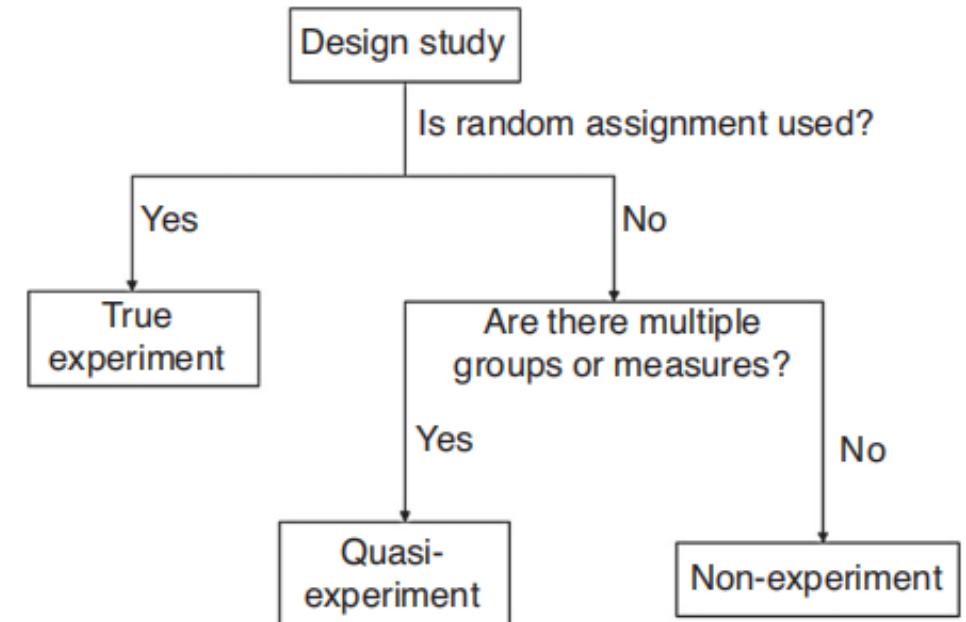


实验设计

- 什么是实验、准实验、非实验

- 真正的实验具有如下特点

- 以至少一个可检验的研究假设为基础，并旨在验证它
- 通常至少有两种条件(实验条件和对照条件)或组(实验组和对照组)
- 因变量通常使用定量测量
- 通过各种统计显著性检验对结果进行分析
- 以消除潜在偏差为目标来设计和进行
- 具备不同的参与者样本，在不同的时间，不同的地点，由不同的参与者进行复现



实验设计

- 大多数成功的实验都是从一个定义明确的、合理范围内的研究假设开始的
 - 研究假设是基于早期探索性研究的结果产生的，并提供了设计实验所需的关键信息——自变量和因变量
- 自变量的数量和值直接决定了实验有多少条件
 - 假设：使用鼠标、操纵杆或轨迹球来选择不同大小(小、中、大)的图标时，目标选择速度没有差别。
- 这里的自变量是谁？取值如何？
- 因变量是谁？如何度量？
 - 可以使用多种方法来测量因变量
 - 可以通过每分钟输入的单词数来衡量，即输入的单词总数除以输入这些单词所用的分钟数
 - 也可以通过每分钟输入的正确单词数来衡量，即输入的正确单词总数除以生成这些单词所用的分钟数

总共有9种($3 \times 3 = 9$)实验条件
实验的目的决定了哪种方法更合适



实验设计

- 如何控制自变量来创造多种实验条件
- 举例



- 假设正在开发一个基于语音的应用程序，并且需要研究识别错误如何影响用户的交互行为
- 有哪些实验条件？
 - 控制条件：语音识别器不会出错，能正确识别用户的每一个单词
 - 对照条件：语音识别器会出现一定比例的错误
 - 问题：所有的语音识别器都会出错
 - 解决方案：绿野仙踪法 Wizard-of-Oz，让一个人充当语音识别器，该方法允许我们测试现实世界中不存在的理想应用程序



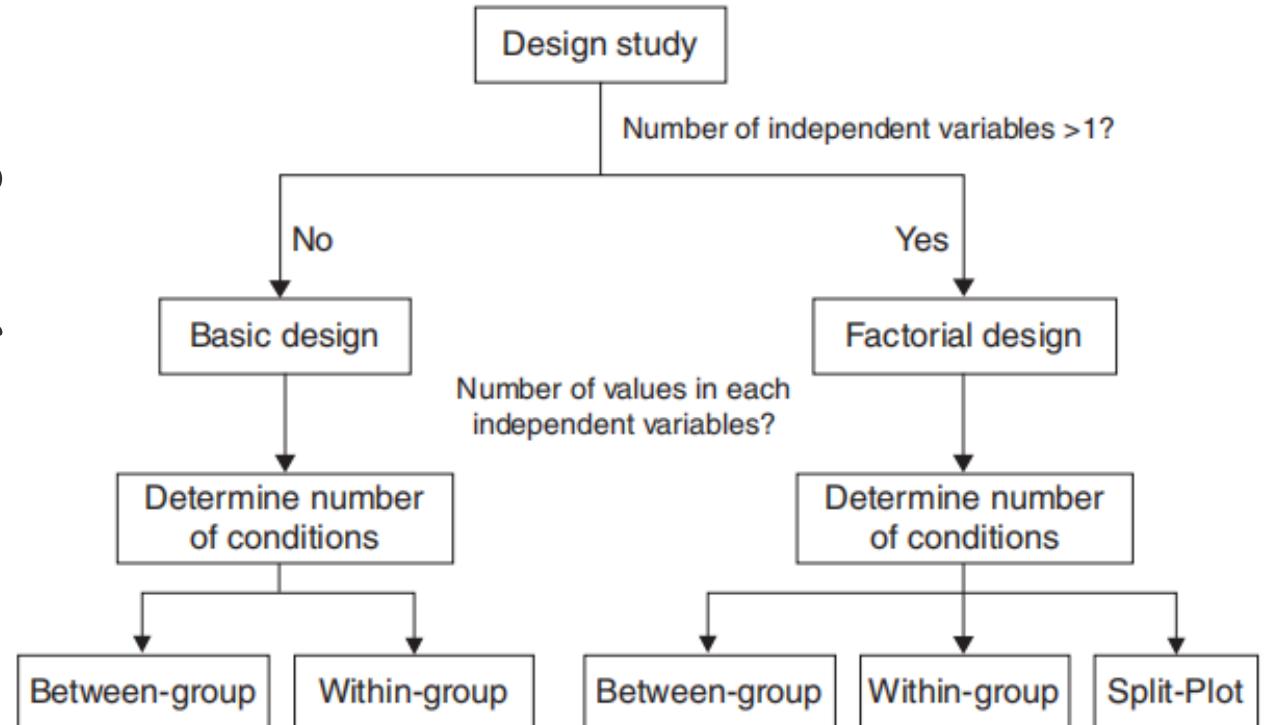
实验结构

- 实验中我们要研究多少自变量?

- 基本设计 or 析因设计

- 每个自变量有多少个不同的值?

- 条件数 => 组间设计、组内设计、裂区设计



一个自变量的实验

- H1:在使用QWERTY键盘、 DVORAK键盘或字母顺序键盘时，打字速度没有区别。
 - H2:在网上商店中，新手(novice)和有经验的用户在找到物品所需的时间上没有差别。
 - H3:来自美国、俄罗斯、中国和尼日利亚的客户对在线代理的感知信任没有差异。
-
- 上述实验中的实验条件是什么？



组间设计和组内设计

- Between-group design (between-subject design)

- 每个参与者只暴露在一种实验条件下
- 参与组的数量直接对应于实验条件的数量

- 优点

- 设计更简洁
- 避免了学习效应

- 缺点

- 结果受个体差异影响大
- 样本量大
 - 如果一个实验有四种条件，每种条件下需要16名参与者，那么需要的参与者总数为64人

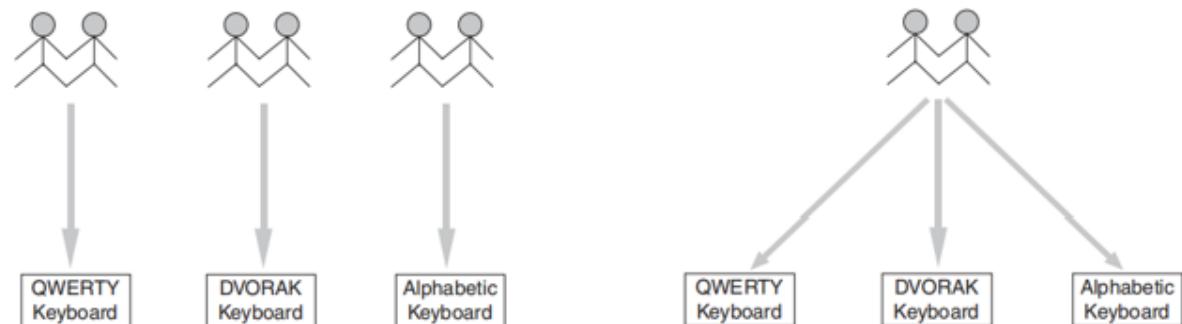


图 3.3 组间设计

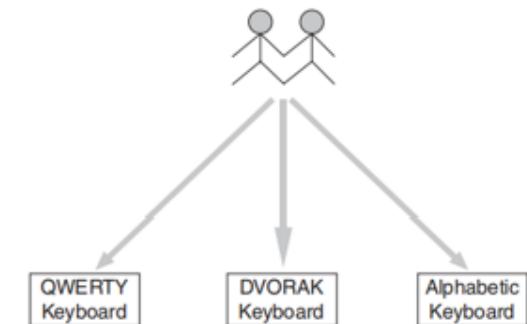


图 3.4 组内设计



组间设计和组内设计

- Within-group design

- 优点

- 样本量小

- 隔离了个体差异

- 缺点

- 学习效果的影响

- 疲劳问题

- 组间设计和组内设计的优缺点完全相反

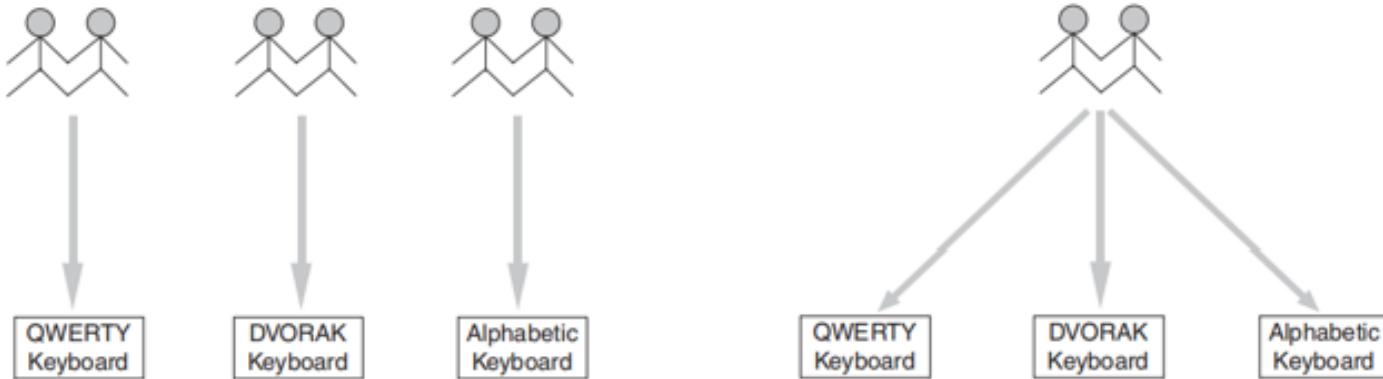


图 3.3 组间设计

图 3.4 组内设计

	Type of experiment design	
	Between-group design	Within-group design
Advantages	Cleaner Avoids learning effect Better control of confounding factors, such as fatigue	Smaller sample size Effective isolation of individual differences More powerful tests
Limitations	Larger sample size Large impact of individual differences Harder to get statistically significant results	Hard to control learning effect Large impact of fatigue



选择合适的设计方法

• 组间设计

- 任务简单，个体差异有限
 - 当任务简单且认知过程有限时，个体差异较小，如在屏幕上选择一个目标的任务
 - 当任务复杂或涉及重大认知功能时，个体差异更大，如阅读、理解、信息检索和解决问题相关的任务
- 受学习效果影响较大的任务
 - 如：在一个比较网站中两种类型菜单导航效果的实验，在一个条件下完成导航任务的参与者会获得大量关于网站架构的知识
 - 在一个条件下完成导航任务的参与者会获得大量关于网站架构的知识
 - 必须采用组间设计



选择合适的设计方法

- 一些不可能采用组内设计的实验

- H2：在网上商店中，新手和有经验的用户在找到物品所需的时间上没有差别。
- H3：来自美国、俄罗斯、中国和尼日利亚的客户对在线代理的感知信任没有差异。
- 问题：一个人不可能同时是新手和有经验的在线商店用户

- 注意：

- 参与者应尽可能随机分配到不同的条件下
- 需要在不同条件下尽量平衡潜在的混杂因素
 - 如性别、年龄、计算机经验和互联网经验。
 - 换句话说，除了作为实验变量的个人特征外，我们需要确保这些群体尽可能地相似。



选择合适的设计方法

- 组内设计

- 个体差异较大、学习效果不太容易受到影响的任务、或目标参与者群体很小的任务
- 需要考虑如何控制与组内设计相关的学习效果、疲劳和其他潜在问题的负面影响
- 如：大多数测试复杂或学习技能或知识的任务——比如打字、阅读、写作和解决问题

- 很难找到和招募合格的参与者是许多HCI研究人员经常面临的问题

- 控制学习效果

- 将实验条件的顺序随机化
- 当研究的目标不是与应用的初始交互时，减少学习效果影响的有效方法是提供充足的培训

- 解决疲劳问题

- 单个实验的适当长度应该是60 - 90分钟或更短，并提供必要的休息机会

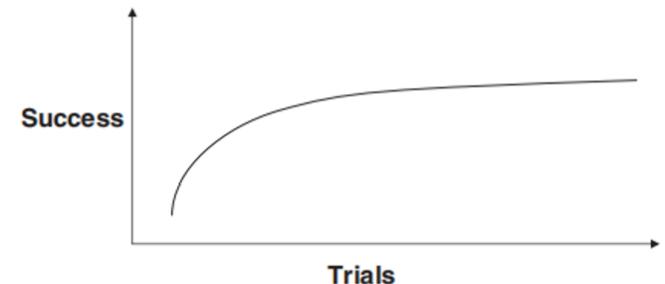


图 3.5 典型学习曲线



选择合适的设计方法

- 学习和疲劳是不可避免

 - 使用拉丁方设计Latin Square进行平衡

- 举例

 - 假设有4种实验条件“ABCD” 和4个测试用户“甲乙丙丁”，如何应用拉丁方设计实验过程，尽量减小学习和疲劳的影响？

4 x 4 Latin Square

A	B	C	D
B	C	D	A
C	D	A	B
D	A	B	C

4 x 4 Balanced Latin Square

A	B	C	D
B	D	A	C
C	C	B	A
D	A	D	B



多个自变量的实验

• 析因设计 Factorial Design

- 当一个实验调查一个以上的自变量或因素时，析因设计被广泛采用
- 可以同时调查所有自变量的影响以及多个变量之间的交互影响
- 析因设计中条件的数量由自变量的总数和每个自变量的取值决定

$$C = \prod_{a=1}^n V_a$$

• 举例

其中 C 为条件个数，V 为每个变量的取值个数。

- 比较使用三种类型键盘(QWERTY键盘、DVORAK键盘和字母键盘)时的打字速度，且调查不同的任务(写作和抄写)对打字速度的影响

- 自变量及取值?

条件数= $3 \times 2 = 6$

- 变量 “键盘类型” 有三个值：QWERTY键盘、DVORAK键盘和字母键盘
- 变量 “任务类型” 有两个值：抄写和写作



多个自变量的实验

• 析因设计 Factorial Design

➤ 分析数据时，比较同一行可以检查不同键盘的影响；任务效果可以通过比较同一列的数据来检验

➤ 组内设计or组间设计？

均可，重要的是平衡实验中的顺序和条件，那么应如何做？

	QWERTY	DVORAK	Alphabetic
Composition	1	2	3
Transcription	4	5	6



裂区设计split-plot design

- 析因研究中的一种设计，既有组间成分，也有组内成分
- 举例
 - 研究问题：年龄和GPS的使用情况
 - 自变量？
 - “年龄” 有三个值：20到40岁的人，41到60岁的人，以及60岁以上
 - 第二个变量有两个值：无GPS驾驶和有GPS辅助驾驶

	20 to 40 years old	41 to 60 years old	Above 60
Driving without GPS assistance	1	2	3
Driving with GPS assistance	4	5	6



多个自变量的实验

- 析因设计 Factorial Design

- 优点：允许在一个实验中研究两个或两个以上自变量之间的相互作用的影响
- 相互作用可被描述为“一个自变量对因变量的不同影响，取决于另一个自变量的特定取值”

- 举例

- 调查设备类型（鼠标和触摸屏）和经验如何影响目标选择任务的有效性
- 两种类型的用户：新手用户和经验用户

- 启发

- 在最初的交互过程中，触摸屏的表现要优于鼠标。但是用户可以在学习鼠标上取得比触摸屏更大的进步，最终使用鼠标达到更高的效率
- 在通常交互时间较短且训练机会有限(如ATM接口)的情况下，触摸屏是一种更合适的输入设备。相比之下，鼠标可能更适合长期、频繁的任务

