测验的开发与题目分析

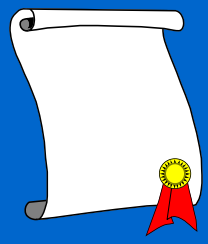
闫巩固 博士

本章要点

- 了解测验开发的一般程序
- 了解题目编写的基本规则
- 题目分析
  - 难度
  - 区分度
  - 题目反应理论

测验开发的一般程序

- 测验界定
- 题目编写
- 题目分析
- 题目与测验的修正
- 测验的出版



第一步：测验的界定

- 测量对象：什么人，儿童或/和成人，
- 测量目标：什么内容，成就-智力-人格---
- 测量目的：诊断/预测
- 测验建构的基础或途径：
  - ♣ 基于逻辑或内容：考虑行为的领域范围并进行相应的度量（如教育测验）
  - ♣ 基于理论：根据一种理论编写出能够反映理论构想的题目（如Myers-Briggs Type Indicator 或MBTI基于容格的类型论）
  - ♣ 基于外部效标：选择那些能够将个体或群体做出区分的题目（如MMPI，CPI，罗夏墨迹测验等）。。基于经验

使用外部效标

- 这是决定删除或保留题目的最流行的一种方法
  - 为什么？
- 例如，加州心理问卷（CPI）中的社会性量表，其题目是那些与一组人中有多少朋友具有高相关的题目
  - 社会性越强→朋友越多

使用外部效标（续）

- 与理论无关(atheoretical)
  - 能够区分各种人群（临床/正常上的子群体）
  - 题目可能没有表面效度，只要他们能预测
  - 可能无法预测一组人相对于另一组人将如何回答一个题目
  - 题目可能因样本变化而失效或不起作用（如抽样偏差，样本太小）
- 出现怪异的题目
  - 题目被用于做预测，独立于它们本身的意思
  - 如MMPI中的题目
- “I am so touchy（易于发火） on some subjects that I can’t talk about them”
  - (True=精神分裂症)
- “I have never had a fit or convulsion”（痉挛或抽搐）
  - (True=抑郁症)

使用外部效标（续）

- 运用理论指导你！
- 外部效标在题目的预测性上提供了有用的资料（如社会性题目应与社会测量评定有相关）
- 理论应当用来生成和改进题目，而不是由它本身来预测的

- 基于理论的题目
- 如EPPS
- I would like to write a great novel or play. (ACH)
- I like to avoid responsibilities and obligations. (Autonomy)
- I like to be the center of attention in a group. exhibition
- I like to be in love with someone of the opposite sex.

双向细目表 (table of specification)

- 一种标明测验所包含的内容和要测定的技能（构想）以及每一个内容、技能的相对重要程度的表格。
- 可能是多向的，不限于双向。
- 要目标明确

内容（主题）

	准备	编题	施测	记分	题目分析
术语	工作分析，关键事件有代表性样本（3）	匹配题；反应心向，多项选择（5）	晕轮效应 主被试关系（2）	记分键 合成分数 机器记分（3）	效标，内部一致性，测验同质性（3）
特殊知识	教育目标的分类（2）	客观题与论述题的长短处（4）	影响测验成绩的因素（3）	主客观题的记分规则（3）	决定题目效度的方法，题目分析的目的（3）
理解	解释制定测验计划的目的（2）	（0）	（0）	题目权重对总分的影响（1）	解释P和D的关系（1）
应用	对某个单元编制一个具体说明（1）	多项选择测量理解、应用、分析、综合和评价能力的例子（4）	测验指导语（2）	猜测的校正，置信加权，（4）	难度和区分度指标计算，错误选项的分布（4）
总计	8	13	7	11	11

第二步：题目的编写

- 最经常讨论的有关自陈测验的问题：

1. 题目选择(item selection)
2. 反应格式(response format)
3. 题目措辞(item wording decision)
4. 量表建构(分量表）(scale construction)
5. 反应心向(response set)
6. 反应风格(response style)

1、选择题目类型

- 题目类型（为什么？测什么）

特点是什么？

- 客观题：
  - 只有一个正确答案
  - 是非题，多选一，匹配题
- 主观题：
  - 许多答案可能都是正确的
  - 填空题，简答题，论述题

用于什么方面的测量？

技能展示

2、题目反应格式

- **是非题（true-false responses）**
  (e.g., Marlowe Crowne Social Desirability Scale)
- **利克特评分量表（Likert-type rating scales）**
- **多重选择题（Multiple-choice items）**
  (for knowledge assessment)

利克特式量表

- **几乎不需要练习**
- ***例题：“Once I find the right way to do something, I stick to it” (NEO)***
- **你会使用哪种格式？**
- **多少刻度点？5，6, 7，10，100…..**
- **使用语言标识？**

3、题目措辞

- 两种主要策略
- **遵循题目编写规则：**
  - **题目长度**（item length）
  - 可读性（readability）
  - 语法（grammar ， e.g., double negatives)
  - 避免男性至上、种族主义以及冒犯性语言 (avoidance of sexist, racist, offensive language)
- **运用统计分析去除坏的题目**

多择一题型的编写规则

- Do…
  - 使用一个直接的问题或一个不完整的陈述作为**题干**
  - 选项或题枝在语法和形式上保持一直
  - 对于那些不知道答案的人，所有错误选项都应看起来是可能的
- Don’t…
  - 否定性题目，尤其是在题干是否定性的（双重否定）
  - 使用技术上的行话
  - 反应之间相互重叠
  - 使用“以上所有”

主观题：人格

- 是非题
  - 我喜欢干有冒险性的工作
- 利克特评定量表：5点或7点评定
  - 朋友们通常把我看成是一个潇洒的人 1 2 3 4 5
    很不象我 非常象我
- 使用特殊符号
  - 用于态度调查
  - 你对XXX电视节目有什么看法？ ☹ ☺

主观态度调查题目的编写规则

- Do…
  - 使用现在时态
  - 避免模糊与歧异
  - 使用简单明确的语言
  - 语句尽量短
  - 选择各种陈述以覆盖所有感兴趣的内容
- Don’t…
  - 询问事实
  - 使用每个人都会赞同或不同意的陈述
  - 使用“如果可能”
  - 使用双重否定
  - 使用全称命题（所有…）

这些题有什么毛病？

- 当上级布置任务时，说不是不能接受的。
- 我不相信我们的所有法律对公民都是有益的。
- 如果你去问的话，人们不会告诉你他们大多数时间都在想什么。
- 有多少人居住在香港？
  - A 超过三百万　B 超过四百万
  - C 超过五百万　D 超过六百万

4、量表建构

- 将题目组织起来
- 与前面一样，三种方法：

1. 推理性量表(rational scale)
   - 需要好的理论（EPPS)
2. 经验性量表(empirical scale)
   - 因素分析（MMPI）
3. 推理－经验性量表
   - NEO－PI

5、反应心向

1. **社会赞许性（Social Desirability）**
2. **随机反应（Random Responding ）**
3. **虚假反应（Dissimulation）**

社会赞许性

- 指按照社会期望的方式反应（概率）
- 问题：区分取悦自己与取悦他人？
- 争论：
  - 是误差变异还是人格的一部分？（如：Agreeableness）
  - 方法学意义？

社会赞许性的控制

- 策略：

1. 避免使用高赞许和非赞许的题目
2. 使用SD量表并运用统计控制赞许性(如，MMPI的K量表， **Marlowe-Crowne赞许性量表**)

随机反应

- 通常表明低的动机（Unwillingness, “ignorance”, just not paying attention ）
- 控制策略：
  - 包括这样一些题目：每个人的反应都倾向于真或假

虚假反应

- **表现为有策略地歪曲或作假（招聘等）**

➡ **不是单一控制而是采用复合证据（Not a single control strategy，Rather combination of evidence）**
➡ **探究动机(Explore motivation of test taker)**
并且
- **有证据表明虚假反应无关大局（Been shown not to matter，some personality tests “robust” against faking, e.g. MMPI)**
- **虚假反应对有些测验不起作用（For some tests shown not to work，如fake instructions)**

6、反应风格

- 指没有好好阅读题目就做出同意或不同意的反应（**Agree or disagree with statements without attending to their actual content, or** ）

➡ **默认反应（Acquiescence， yeah-saying, tend to agree after not reading the question)**
➡ **挑剔反应（Criticalness， nay-saying, tend to disagree)**

➡ **可能对模棱两可的题目这是一个严重问题。**
➡ **采用反向计分题**

一些考虑与区分

- 不管是主观题还是客观题，命题时都需要考虑
  - 内容----过程
  - 速度----难度

可能影响人们回答的其他因素

- 问题呈现的顺序
- 问题出现的情境
- 问题是开发的还是封闭的
- 问题是否经过了过滤
- 问题是否包含某些时髦词语
- 答案选项的范围
- 答案选项出现的顺序
- 是否提高了中间选项
- 问题是从受益还是损失的角度来提问的

斯科特.普劳斯：《决策与判断》，p67

小结：题目编写

- 应用理论指导你的题目编写，不是只靠与外部效标的相关
- 简明扼要-明确的指导语，简明的语言
- 校对你的题目
- 对可能的混淆有所预期
- 考虑疲劳/厌倦因素
- 考虑短时记忆的限制

第三步：题目分析

- 测验由题目构成。题目的质量决定了测验的质量。
- 题目分析可以帮助我们发现题目的问题，并加以调整。在测验建构中起着重要作用。
- 题目分析可以从质和量两个方面进行。
  - 质：内容与形式
  - 量：统计学特性

题目分析的两个基本方面

- 难度(difficulty)
  - 人们做对一个题目的比例
- 区分度(discrimination)
  - 题目是否对不同类型的人群有不同的难度（如高分者和低分者；性别或种族等各种人群）

题目难度

- 定义：正确作对一个题目的比率或通过百分比(percentage passing)
  - P=作对人数/总人数 （0 或 1 记分）
  - P=平均值/最大值 （连续记分）
- P的相对性
  - 因样本而变化

最佳难度

- 最佳难度的确定
  - 设定测验的通过率
  - 如，你想录取30%的人，那么，应选择那些题目，其平均正确率为30%
  - 一般而言，一个测验的P=.50是最佳的。WHY？
- 当P=.50时，分数（0-1）的标准差最大
  - SD=（P（1-P））$^{1/2}$
- 为什么不要P=1或P=0的题目？
  - 因为他们没有为区分人提供有用信息。

其他问题

- 测验分数的分布
  - 正偏态（平均值位于低分一端）
  - 负偏态（平均值位于高分一端）
  - 当标准样本的分数分布为偏态时，需要对题目进行调整。它可能包含了很多过难或过易的题目。增加或删除一些题目。
- 等距量表
  - 由于P是百分数，为了便于比较，我们可以将其转换为平均值为13，标准差为4的标准分数
- 不同年级学生的比较（瑟斯顿绝对量表，Thurstone absolute scaling, 1925,1947）

区分度

- 题目的区分度是指一个题目正确区分测验想测行为的程度。这与效标效度有关。也称为题目效度。
- 题目区分度的统计指标
  - 在建构测验上，至少目前已发展出50多种不同的题目区分度统计指标
  - 尽管这些指标是不同的，但绝大多数的值是相近的

区分度指标

- 相关指标
  - 测验分数与题目分数的相关
- 积差相关
- 二列相关
- 点二列相关
- 四分（Phi）相关
- 。。。

- 鉴别力指数
- $D=P_H-P_L$
  - P_H=高分组的通过率
  - P_L =低分组的通过率
- 分组可以是测验总分，也可以是效标分数
- D 在-1和+1 之间

极端组的确定

- 27%规则（T. L. Kelley,1939)
- 方便方法：25%-33%
- 难度与区分度
  - 如果一个测验的每个题目都是P=.50，那么，它会是一个好测验吗？

例：四个题目的难度与区分度

题号	组别	a*	b	c	d	空白	p	Δ	r_b	D
1	高	36	0	39	23	2	.34	14.7	.04	.04
	低	32	0	46	18	4				
2	高	22	12	10	48	8	.27	15.6	-.12	-.10
	低	32	25	11	23	9				
3	高	62	16	7	15	0	.44	13.6	.37	.36
	低	26	36	7	28	3				
4	高	95	2	1	2	0	.78	10.0	.55	.41
	低	54	18	12	16	0				

a为正确选项

郑日昌：209

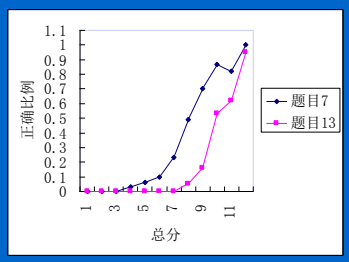
鉴别力评价

- 伊贝尔（L.Ebel)
- 鉴别力D　　评价
- .40以上　　非常优良
- .30-.39　　良好，如能改进更好
- .20-.29　　尚可，需要修改
- .19以下　　差，淘汰或改进

题目反应理论（Item Response Theory, IRT）

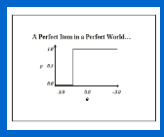
- 题目特征曲线(item characteristic curve, ICC)描述了潜在能力与作对一个题目的概率之间的关系。
- ICC可以反映了题目的特性
  - 难度，区分度与猜测

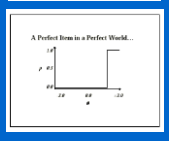
例子



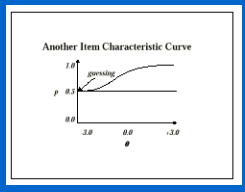
IRT的基本假设

- 假如有一个低能力的人，那么，在测量是理想状况下，他应当只能作对很少的题目
- 假如有一个高能力的人，那么，在测量是理想状况下，他应当能作对更 多的题目

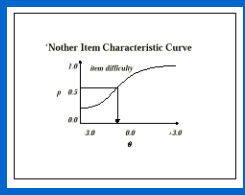




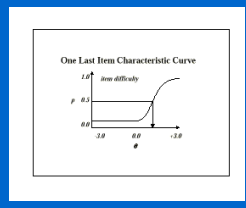
- 事实上，现实世界并不存在完美的度量
- 人们在做题时，尤其是难题时会依靠猜测。
- 对于一个四择一选择题，人们是靠猜测作对的概率为25%



- 心理能力可采用心理物理学中阈限测量的概念，可定义为50%答对的那一点



- 区分度可定义为ICC的斜率

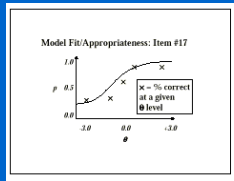


ICC

- 难度：在ICC上对应于能力的点
- 区分度：ICC的斜率
- 猜测：在ICC上最小的P值

我们如何应用这些思想

- 需要大量的被试来估计题目的参数，即难度，区分度和猜测
- 每一个ICC模型与被试的实际反应相拟合



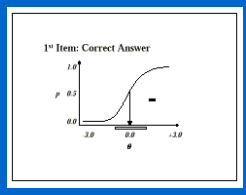
IRT的实际用途

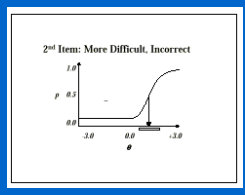
- 计算机适应性测验(computer adaptive test, CAT)
- 题目偏差(Test bias)
- 题目等值(test equating)：创造平行测验（复本）

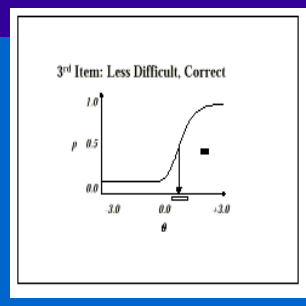
计算机适应性测验

- 基本原理：根据测验者的反应，由计算机自动裁决下一个题目的出现
- 给出一个中等难度的题目
  - 答对：出现一个更难的题目
  - 答错：出现一个更易的题目
  - 循环往复，直至得到在一定精度上的某个能力值（θ）

例子







计算机适应性测验的例子

Example: Estimating a Person's θ (Verbal Ability)
Prior θ = 0.00, SE(θ) = 1.00
Final θ = 1.23, SE(θ) = 0.19

Item Presented	*Item Number*	*Item Response*	*Ability (θ) Estimate*	*Standard Error*
1	43	+	.469	.86
2	257	–	.328	.75
…	…	…	…	…
40	211	+	1.23	.23
41	148	–	1.22	.21
42	92	+	**1.23**	**.19**

CAT的好处

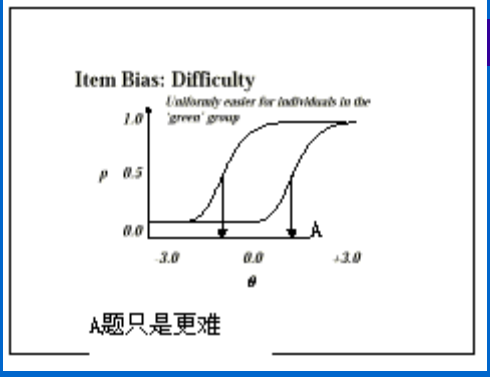
- 不需要做所有测验题目就可以获得更加精确的能力估计值
- 减少疲劳，测验时间，以及被试的挫折感等

CAT的劣势

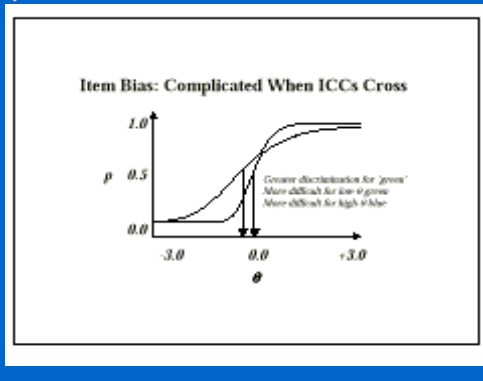
- 开发程序需要大量金钱
- 良好的估计需要大量题目（题库）和被试
- 在以下情况下，人们感觉有不公平
  - 不同的人做的是不同的题目
  - 不同的人做测验的时间不同
  - 测验题目的加权值不同
- 但一般认为，CAT更公平

题目偏差

- 一个题目对不同样本可能有不同的特征（性别，民族，年级组）
- 一个题目对不同组的人可能有不同的难度、区分度和猜测率



A题只是更难



检查题目偏差的其他方法

- 黄金规则(golden rule)
- 根本不考虑能力值
- 将不同组之间P值的差异大于15%的题目删除
- 存在的问题
- 有些组之间确实存在真正的差异
- 没有组间的差异有时很难编制“平行”（等价）的题目

Mantel-Haentszel 技术

- 按测验总分将不同组的成员进行等值匹配
- 计算不同组在题目上的通过率
- 通过率有差异的题目界定为存在偏差
- 问题是测验总分可能不是能力的最佳度量

Example: Mantel-Haentszel Analysis

(for each group, *p*-values by score level)

	p-values for an item		
	Group A	Group B	Group C
High scores	.20	.40	.50
Middle scorers	.30	.40	.30
Low scoreers	.50	.20	.20

第四步：测验修改

- 根据项目分析结果，删除、修改和重新编制题目。这是一个多次往复的过程
- 对成型的测验进行标准化，收集测量学证据或资料，包括信度与效度等

第五步，测验出版

- 这一步是可选择的
- 测验材料
- 用户手册
- 技术手册
- 出版商

最后的问题

- 所有题目测量同一种东西吗？
- 如果测量两次，会得到相同或近似的结果吗？
- 如果两个评分者改卷，得到的结果相同吗？
- 测验内容适当地测量了我们想要测的东西吗？
- 我们得到的结果适当地反映了要测的东西吗？
- 我们实际测量的是我们想要测量的东西吗？