**Psychometrics index，基于CTT（经典测量理论）**

**宏观指标梳理/目录**

**难度(项目分析)/单项题目能否有效得出结果，仅用于能力题**

**区分度(项目分析)/单项题目能否有效区分目标人群，仅用于能力题**

**信度/可靠性：稳定性、复本和一致性**

**效度/有效性**

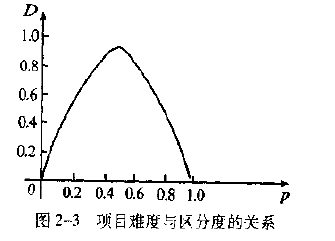
**难度和效度可以被包装成准确性。**

**Spearman-Brown formula可以用来解决试验项目长度改变的问题。**

**IRT与CAT可以用来解决试验整体长度缩短的问题，通过扩大题库和缩减答题时间**

**项目分析：**

**\* 难度0.5，区分度1的项目为最佳。其曲线如下：**



* 难度（具有正确选项的题型）：

\* 分数在前后27%~30%之间的被试为高低分组；假设75个人30%是22.5个人，27%是20.5个人，那么取21、22个人都没什么问题，综合观察样本量决定。

\* 0.3~0.7均为合格区间

总通过率：

极端分组通过率：

|  |  |
| --- | --- |
| 指标 | 解释 |
| P | 通过率 |
| R | 答对或通过该项目的人数 |
| N | 全体被试人数 |

* 难度校正公式（选择题）：

|  |  |
| --- | --- |
| 指标 | 解释 |
| CP | 校正通过率 |
| KP | 实得通过率 |
| K | 备选答案数目 |

* 难度（具有量纲的题目，一般不适用，不建议）

总通过率：

极端分组通过率：

|  |  |
| --- | --- |
| 指标 | 解释 |
| P | 通过率 |
| Mean | 项目均值 |
| Scale | 项目测评尺度，如Likert5点，则为5 |

* 区分度（具有正确选项的题型）：
  + 鉴别指数：单项目（常用）

D越大，项目区分度越大，则越有效。D>0.4时，项目为很好；0.39>D>0.30时，项目为良好，需要修订；0.29>D>0.20时，项目为尚可，仍需修订；D<0.19时，项目应该被淘汰。

* + 二列相关/点二列相关：总分、具有校标效度。此处只列出二列相关

当校标总分近似为连续（此处为了严谨，实际使用时，正确与否的题目保证一定题目量即可）

|  |  |
| --- | --- |
| 指标 | 解释 |
| r | 相关性 |
| Xp | 与二分变量通过组对应的连续变量的平均数 |
| Xq | 与二分变量未通过组对应的连续变量的平均数 |
| p | 通过组人数与总人数之比 |
| q | 未通过组人数与总人数之比 |
| S | 连续变量的标准差 |
| y | p与q交界处正态曲线的高度 |

具有显著性公式，此处略

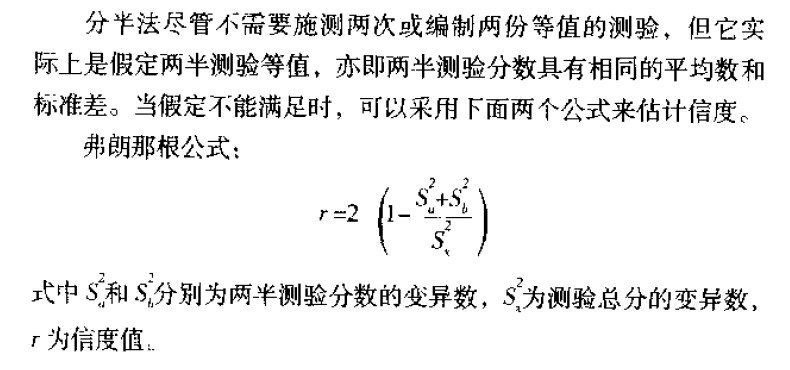
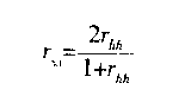
* 区分度（具有量纲的题目，一般不适用，不建议）：

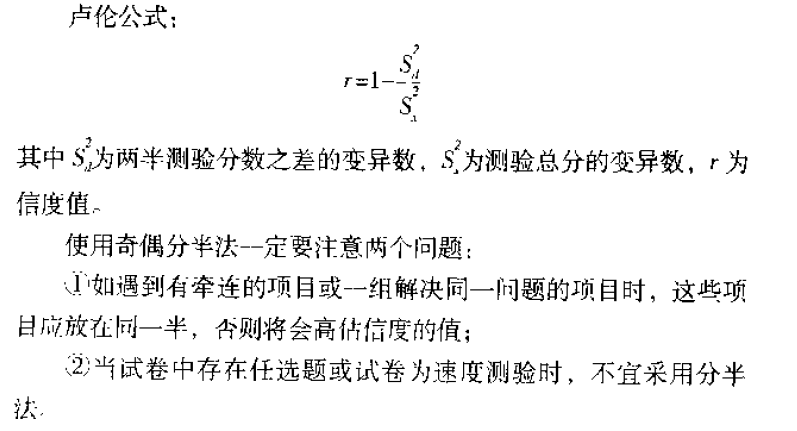
项目与总分相关系数

**传统信度：**

Top3：stability重测, alternate form复本（包含分半，严格复本与校标相似）, and consistency一致性

* 重测信度（test-retest reliability）
  + 定义：用同个试验对同一组被试进行前后测。时间需要拉到一个评估练习效应消失的足够阈值。
  + 测量指标：ICC组内一致性系数、相关系数r（简便做法）、Kappa、Kendall W等
* 复本信度（alternative-form reliability/parallel forms reliability）：可用于GIA&PI的计算
  + 定义：两个等值测验上得分系数的相关程度
  + 测量指标：相关系数r
* 分半信度（split-half reliability）
  + 定义：将测验拆分成一半（随机抽样，考虑多个因素，如奇偶）。需要使用Spearman-Brown formula进行校正
  + 暂时先截图：





* + 测量指标：
* 内部一致性信度/同质信度（最常用）
  + 定义：测量同一个概念的多个计量指标/测量题目/items的一致性程度
  + 测量指标（最常用）：Cronbach’s α（主观题）/KR20（能力题）
* 评分者信度（不常用）/评分者一致性

**传统/过时效度：**

* 专家效度
  + 实施方法：邀请相关领域的专家对结果进行判定
  + 缺陷：没有数据驱动，可信性不高，容易受怀疑。另外，设计方案也需要定义。仅用于没有办法的办法。
  + 测量指标：一致性系数（相关）等、定性评价等
* ~~内容效度：教育测验/定性程度较强的测验。（本质上使用心理测量学的优质流程+专家效度）~~
  + ~~实施方法：确定所要测量的全部内容范围；将测验目的具体化为不同层次的测验目标；确定每一层测验目标在整个测验中的比重；编制双向细目表；确定每一小格中的测题数量；按比例随机抽取测验题目；请专家对测题的代表性、适合性进行分析，并作必要的修改，直到大多数专家满意为止。~~
* 探索性因素分析（EFA）仅用于未知题库，未知因素的情况下使用，数据驱动、效率不高、缺乏定性文档。
* 同时效度/并行效度（Concurrent validity）/效标效度（Criterion validity）
  + 定义：与现有标准（criterion）之间的对照，在一定程度上可包含专家效度
  + 测量指标：一致性系数（相关）等

**流行效度（均为数据）：**

* **验证性因素分析（CFA）：**
  + **结构效度/构想效度（construct validity）**：一般CFA相关指标，有些测量误用EFA进行报告（根据测量方式而定）
  + **组合信度CR（composite reliability）：**
    - 源于经典测量理论
  + **区别/区分效度（discriminant validity）：**
    - 定义：不同构念的题项仅存在适当的相关。
    - 测量指标：观察loading值，AVE的平方根大于这个变量与所有其他变量的相关系数
  + **聚合效度（convergent validity）：**
    - 定义：聚合效度指的是相同构念的题项能够存在很强的相关，
    - 测量指标：cronbach’s alpha > 0.7，AVE > 0.5
* Campbell和Fiske（1959）曾经开发出一套经典的方法：多元特征与多重方法矩阵（Multitrait-Multimethod Matrix，MTMM）【多质多法模型】来同时研究这两者。
* **预测效度Predictive validity：**
  + 根据预测效标材料确定的有效性,反映的是从人才测评测验分数预测任何效标情境或者 一段时间间隔后被测者行为表现的程度
  + 测量指标：需要专业设计，通常与测量目的有关。

Convergent validy 用来检验潜变量（即construct）对应的观测变量（即item）是否收敛于该潜变量，discriminant validiy用来检验潜变量对应的观测变量组与其他观测变量组之间是否有明显区别。由于潜变量与其观测变量之间的关系有所谓reflective 和 formative两种情况，平常我们用到的大多是reflective 情况。对于reflective construct，其完整的Convergent validy 检验应该包括：1、各item loading大于0.7,AVE>0.5，OK; 2、item loading  
在04.-0.7之间，如果删除对应item，AVE和CR值增加超出0.5和0.7的阈值，该item就删除，否则保留；3、item loading<0.4，毫不犹豫删除。 完整的discriminant validity检验包括：1、在item相关关系矩阵中，item loading> cross loading；2、符合Fornell-larker 法则：潜变量的AVE平方根值大于该潜变量与其他潜变量之间的相关系数。Formative construct 检验的情况就不说了。 以上源自Hair等2013年出版的一本新书，作者都是SEM和多元统计分析领域的绝对权威。