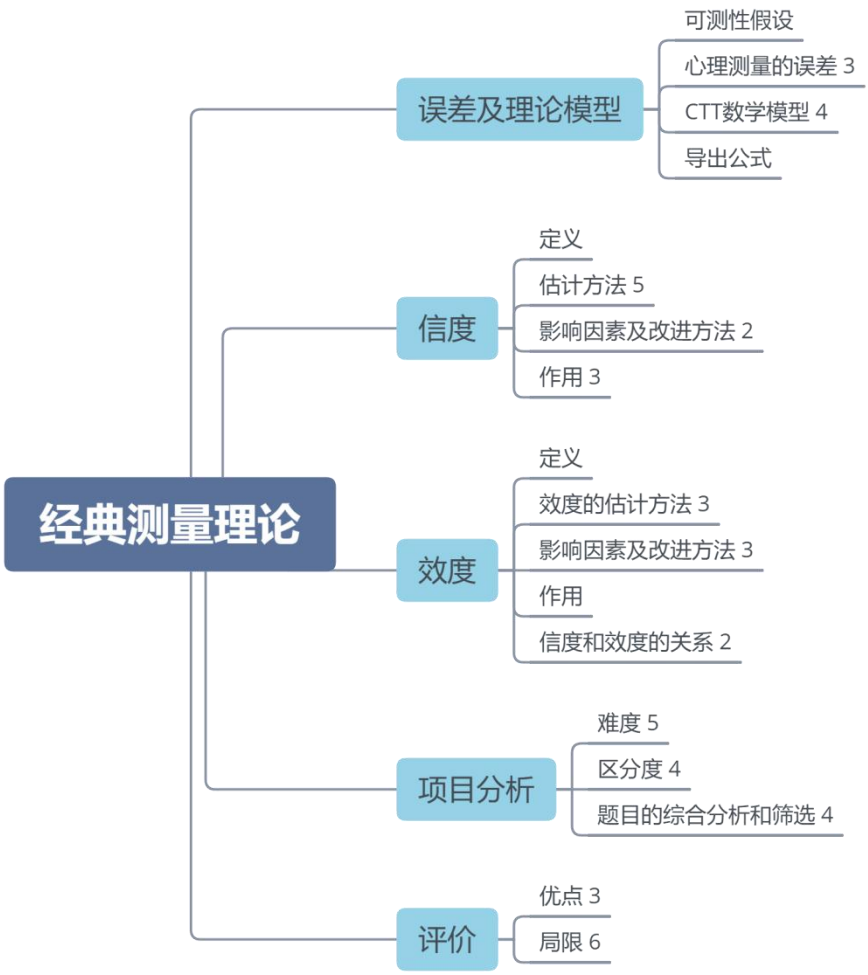
第二章　经典测量理论



一、测量误差及理论模型

（一）可测性假设，心理特质可测量需要两个假设：

①凡客观存在的事物都有其数量。（桑代克，E. L. Thorndike）

②凡有数量的东西都可以测量。（麦克考，W. A. McCall）

心理特质是一种客观存在，具有以下特征：

1.特质是一组具有内部相关的行为的概括，具有一定的抽象性。

2.特质是“一种一般的神经心理系统，它可以综合不同的刺激，使人对这些刺激做出相同的反应”。

3.特质是一个人身上比较稳定的特点。

4.一个人的精神面貌是由多种特质分多个层次有机组合而成的。

5.特质可以决定一个人对特定刺激的反应倾向，可以对人的行为进行某种预测。

（二）心理测量的误差

1.含义

测量误差是指在测量过程中由那些与测量目的无关的变化因素所产生的一种不准确或

不一致的测量效应。它的含义包括：

①测量误差是由那些与测量目的无关的变因所致；

②测量误差表现为不准确（效度低）或不一致（信度低）两种方式。

2.分类

（1）随机误差

随机误差是由与测量目的无关的偶然因素引起的不易控制的误差。它使多次测量产生了不一致的结果，其方向和大小的变化完全是随机的，只符合某种统计规律。又称观察误差、偶然误差、测量误差。

（2）系统误差

系统误差是由与测量目的无关的变因引起的一种恒定而有规律的效应。这种误差稳定地存在于每一次测量之中，此时尽管多次测量结果非常一致，但实测结果仍与真实数字有所差异。

系统误差只影响测量的准确性（即效度），而随机误差既影响准确性（即效度）又影响一致性（即信度）。

3.来源及控制

（1）来源

①测量工具方面：主要原因是心理测量量表是否稳定、是否真正测到我们所要测的东西。同时，题目过少或取样缺乏代表性、题目格式不妥、难度过高或过低、用词不当、时限过短、测验问题带有欺骗性等都会带来测量误差。

②被测对象方面：主要原因是受测者的真正水平是否得到正常发挥。受测者的某种心理特质水平是相对稳定的，但是他在接受测量时的生理和心理状态会影响其水平的正常发挥。同时，受测者应试动机的强弱、受训时间的长短、受训内容的多少、答题反应的快慢、测验经验等都会产生测量误差。

③施测过程方面：主要原因是一些偶然因素，包括施测物理环境，主试的某些属性（如年龄、性别、外表），评分计分环节出现的疏漏，以及意外干扰等。

（2）控制

①测量工具方面：提高编制测验的科学性。要注意搜集材料的丰富性和普遍性，同时还应注意项目取样的代表性、项目难度有一定的分布范围、测验用语简单明了等。

②被测对象方面：主试和被试相互配合及规范操作。

③施测过程方面：所有受测者须在相同条件下接受测试，评分要具客观性，对测验结果的解释要标准化。

（三）CTT 数学模型

1.基本概念

（1）观察分数（X）：实测分数。

（2）真分数（T）：反映被试某种心理特质真正水平的数值，操作定义是无数次测量的平均值。当观察分数接近真分数时，就说这次测量的误差较小。真分数是一个理论上构想出来的抽象概念。

由于误差的存在，在实际测量中真分数是很难得到的。唯一的办法只能通过改进测量工

具、完善操作方法等办法来使观察值尽量接近真分数。只要观察分数和真分数之间的误差不是太大或是被控制在可接受的范围内，测量即可接受。

（3）误差分数（E）：观察分数和真分数之间的差距，即随机误差。

2.模型

X=T+E（经典测量理论假设：观察分数与真分数之间是一种线性关系，并且只相差一个随机误差）。

3.假设公理

（1）若一个人的某种心理特质可以用平行的测验反复测量足够多次，则其观察分数的平均值就会接近于真分数。即：ε（X）=T 或者 ε（E）=0。

（2）真分数和误差分数之间的相关为零，即：ρ（T，E）=0。

（3）各平行测量上的误差分数之间相关为零，即：ρ（E1，E2）=0。平行测验指的是两个题目不同的测验，但它们测的都是同一特质，并且题目形式、数量、难度、区分度以及测验的分布都是一致的。

第一条假设意在说明 E 是个服从均值为零的正态分布的随机变量；第二、三条假设则在于说明 E 是个随机误差，没有包含系统误差在内。

4.说明

（1）在问题的研究范围之内，反映个体某种心理特质水平的真分数假定是不会变的，测量的任务就是估计这一真分数的大小。

（2）观察分数被假定等于真分数与误差分数之和，即假定观察分数与真分数之间是线性关系，而不是其他关系。

（3）测量误差是完全随机的，并服从均值为零的正态分布。它不仅独立于所测特质的真分数，而且独立于所测特质以外的其他任何变量，这就保证了 E 中不含有系统误差成分。此外，各平行测验上误差分数间的相互独立也进一步保证了 E 的随机性，使得观察分数的均值可以稳定地趋于真分数。

（四）导出公式

$$S_X^2 = S_T^2 + S_E^2 = S_V^2 + S_I^2 + S_E^2$$

对于一个团体来说，实得分数、真分数和测量误差之间的关系为：$S_X^2 = S_T^2 + S_E^2$ 。这里的误差均是指随机误差的变异，系统误差的变异包含在真分数的变异中，即真分数的变异可以分成两个部分：与测验目的有关的变异 S_V^2（有效的变异数）和与测验目的无关的变异 S_I^2（无效的变异数，即系统误差的变异），即 $S_T^2 = S_V^2 + S_I^2$。

二、测量的信度

（一）定义

信度即是测量结果的稳定性程度、一致性程度，也叫测量的可靠性。信度有三种等价定义：

1.信度是真分数变异与观察分数变异之比：$r_{xx} = S_T^2/S_X^2 = (S_X^2 - S_E^2)/S_X^2 = 1 - S_E^2/S_X^2$

2.信度是真分数与观察分数的相关系数的平方：$r_{xx} = \rho_{XT}^2$

3.信度是两个平行测验间的相关系数：$r_{xx} = \rho_{xx'}$

其中，r_{xx}就是信度，又称信度系数。一般性能良好的能力与学习成就测验的信度系数应达到 0.90 以上，性格、兴趣、价值观等人格测验的信度系数应达到 0.80 以上。

信度指数（ρ_{XT}，信度系数的平方根）描述测量结果的一致性程度。

（二）估计方法

1.重测信度

（1）含义：用同一个测验，对同一组被试前后两次施测所得结果的一致性程度，又称稳定性系数。其大小等于两次测验分数之间的相关系数。估计测验跨时间的一致性。

（2）计算：皮尔逊积差相关

$$r=\frac{\sum xy}{N_{S_X S_Y}}=\frac{\sum xy}{\sqrt{\sum x^2\times\sum y^2}},\ x=X-\overline{X},\ y=Y-\overline{Y}$$

（3）条件

①所测特质需稳定。

②遗忘和练习的效果基本上相互抵消（智力测验间隔 6 个月左右）。

③两次施测之间，被试在所要测查的心理特质方面没有获得更多的学习和训练。

（4）误差来源：主要由时间间隔造成，还包括其间测验条件和受测者身心状况的改变、测验所测特质本身不稳定等。

（5）评价

①优点：能提供测验结果是否随时间而改变的资料，可作为预测受测者将来行为表现的依据。

②缺点：易受练习和记忆的影响，只适用于测量那些不会随时间变化而改变的特质。

2.复本信度

（1）含义：用两个平行的测验，对同一组被试进行施测所得结果的一致性程度。若两个复本是同时连续施测的，称其为等值性系数；两个复本是相距一段时间分别施测的，称其为稳定性与等值性系数（这是对信度最严格的检验，其值最低）。

复本信度的大小等于两个复本测验分数之间的相关系数。等值性系数估计测验跨形式的一致性，稳定性与等值性系数估计测验跨时间和形式的一致性 。

（2）计算：皮尔逊积差相关

（3）条件

①能够构造出两份及以上的真正的平行测验（内容、形式、难易等方面相同或相似）。

②被试要有条件接受两个测验。

（4）误差来源

①对于等值性系数而言，主要是由题目内容造成的，另外还包括被试方面的情形波动、动机变化等；

②对于稳定性与等值性系数而言，除题目内容的影响外，还会受到时间间隔的影响，所以信度较稳定性系数、等值性系数要低。

（5）评价

①优点：应用范围较重测信度的范围大。

②缺点：严格的平行测验很难构造，容易受练习、记忆和迁移的影响，测验的难度会由于重复而有所改变。

3.分半信度

（1）含义

将一个测验分成对等的两半后，所有被试在这两半测验上所得分数的一致性程度。估计跨两个分半测验间的一致性。一般按题号的奇偶性、题目难度、题目内容分半。

（2）计算：同样是计算两半分数之间的积差相关系数，但因为这只是半个测验的信度，还必须使用矫正公式矫正。

①斯皮尔曼一布朗公式

$r_{xx}=\frac{2r_{hh}}{1+r_{hh}}$，$r_{hh}$为两半测验分数间的相关系数，$r_{xx}$为整个测验的信度值。

$r_{nn}=\frac{nr_{11}}{1+(n-1)r_{11}}$，$r_{11}$为为原测验的信度值，$r_{nn}$为测验长度增加为n倍后的测验信度值。

这一只有当两半测验的变异数相等即方差齐性时才能使用，否则应用下列公式。

②弗朗那根公式：

$r_{xx}=2(1-\frac{S_a^2+S_b^2}{S_x^2})$，$S_a^2$和$S_b^2$分别是两半测验的方差，$S_x^2$是测验总分方差。

③卢仑公式：

$$r_{xx}=1-\frac{S_d^2}{S_x^2}，S_d^2是两半测验分数之差的方差，S_x^2是测验总分方差。$$

（3）条件

①通常在只能施测一次或没有复本的情况下使用。

②测验无法分半时不能用。

（4）误差来源：主要来源于题目本身，与时间因素无关。

（5）评价

①优点：可在没有复本的条件下使用。

②缺点：有些题目难以分半，不同分半方法之间有差异，不适合用于速度测验。

4.同质性信度

（1）含义：

指测验内部所有题目间的一致性，也称内部一致性系数。包括两层含义：

①所有题目测的都是同一种心理特质；

②所有题目得分之间都具有较强的正相关。

估计测验跨项目的一致性。

（2）计算

①克隆巴赫α系数：

$$\alpha=\frac{K}{K-1}(1-\frac{\sum S_i^2}{S_x^2})$$

K：题目数，S_i^2：每题方差，S_x^2：总分方差。

②库德里查逊公式 20（K-R_{20}）：仅适合（0，1）计分。

$$r_{xx}=\frac{K}{K-1}(1-\frac{\sum p_iq_i}{S_x^2})$$

③库德里查逊公式 21（K-R_{21}）：

$$r_{xx}=\frac{K}{K-1}(1-\frac{K\bar{p}\bar{q}}{S_x^2})$$

$\bar{p}$ 与 $\bar{q}$ 分别表示平均通过率和平均失败率，要求所有题目难度接近。

④荷伊特信度：无明确使用范围，使用方差分量比描写测验内部一致性。

⑤因素分析。

（3）条件：不是所有的测验都要求较高的同质性信度，这取决于测量的目的。

一般用于预测的测验或学绩测验可以不考虑同质性，但在提出或验证某种心理学理论的概念和假设时，则需考虑。

（4）误差来源：主要来源于内容取样和所研究行为的异质性。

（5）评价

①优点：只施测一次，可以排除练习和记忆的效果。

②缺点：只可在测量单一概念的测验上使用，不适合应用于速度测验。

5.评分者信度

（1）含义

多个评分者给同一批人的答卷进行评分的一致性程度。估计测验跨评分者的一致性。

（2）计算

①评分者人数为 2 个时，使用积差相关或等级相关。

②评分者人数多于 2 个时，使用肯德尔和谐系数（W 系数）。

③被评对象人数多于 7 个时，使用卡方检验χ^2=K（N-1）W ，df=N-1。

（3）条件：适用于一些无法完全客观计分的测验，如创造力测验及投射测验等。

（4）误差来源：评分者之间的差异。

（5）评价

①优点：适合无法客观计分的测验。

②缺点：容易受到评分者主观判断的影响。

（三）影响因素及改进方法

1.影响因素

（1）主试方面

①施测者不按规定施测，故意制造紧张气氛，或给考生一定的暗示、协助等，测量信度会大大降低；

②评分者评分标准掌握不一，也会降低信度。

（2）被试方面

①对于个体而言，被试的心理特质水平的稳定性，如身心状况、注意力、态度等会影响信度；

②对于团体而言，离散程度和团体的平均水平会影响信度。团体同质性越大，全体得分分布越窄，信度越小，可能低估信度，反之，可能高估信度；另外，团体均分过高或过低，

都会使得分分布变窄，而低估真正信度。

（3）测量工具方面

测量工具是否稳定、试题取样、试题难度、测验长度、试题之间的同质性程度（一套测验中同质性题目越多，同一特质被考查的次数越多，全体得分分布越广，信度越高）等会影响信度。

（4）施测过程方面

考场是否安静、光线和通风是否良好、设备是否齐全、桌面是否合乎要求；另外，对于计算稳定性系数及稳定性与等值性系数时，两次测验间隔的时间越短，信度越高。

2.改进方法

（1）主试方面

主试严格执行施测规程，评分者要严格按标准给分。

（2）被试方面

选取恰当的被试团体，提高测验在各同质性较强的亚团体上的信度。

（3）测量工具方面

精心编制测验量表，避免出现较大的系统误差。适当增加测验长度，但新增项目须与原题同质，且新增项目须适度；使所有题目难度接近正态分布，并控制在中等水平，这样得分分布更广；努力提高试题的区分度。

（4）施测过程方面

施测场地按测验手册的要求进行布置，减少无关因素的干扰。

（四）作用

1.信度是测量过程中随机误差大小的反映。

信度很低，随机误差就很大，这种偏差完全是随机决定。测量中的系统误差与信度无关。

2.信度可以用来解释个人测验分数的意义。

$$SE = S_x\sqrt{1 - r_{xx'}},\ X - Z \times SE \le T \le X + Z \times SE$$

SE 为测量误差分布的标准差，又称标准误；S_x 为测验分数的标准差；rxx′为信度系数；X 为观察分数；Z 为某个统计检验显著性水平的标准正态分布下的临界值；T 为真实分数。

3.信度可以帮助进行不同测验分数的比较。

$$SE = S\sqrt{2 - r_{XX} - r_{YY}}$$

S 为相同尺度的标准分数的标准差。r_{XX} 和 r_{YY} 分别是两个测验的信度系数。

先将原始分数转换成相同尺度的标准分数（T 分数、Z 分数），再将标准分数的差异与 1.96SE（0.05 水平）进行比较，即可得出两个测验的差异是否显著。

三、测量的效度

（一）定义

效度是指一个测验或量表实际能测出其所要测的心理特质的程度，等于一组测量分数中与测量目的有关的变异与实得变异之比。

$$r_{xy}^2 = \frac{S_V^2}{S_X^2},\ r_{xy}^2\text{为效度，}r_{xy}\text{为效度系数。}$$

关于效度的概念，需注意：

1.效度是一个相对的概念，是相对一定的测量目的而言的，且测量只能达到某种程度上的准确。

2.效度是测量的随机误差和系统误差的综合反映。

3.判断一个测量是否有效要从多方面搜集证据。

（二）效度的估计方法

1.内容效度

（1）含义：一个测验实际测到的内容与所要测量的内容之间的吻合程度。

（2）用途：适合成就测验、职业测验（选拔和分类)，不适合能力倾向测验和人格测验。

（3）确定方法

①逻辑分析法：又称专家评定法，包括明确范围、编制双向细目表、制定评定量表三个步骤。

②统计法

A.克龙巴赫法：测量同一内容的两套平行测验分数之间的相关，若相关高，则可能有较高的内容效度；若相关低，则至少有一套测验缺乏内容效度。

B.再测法：在学习某种知识前后参加同一个测验，若后测成绩显著优于前测成绩，则明有较高的内容效度。

C.内容效度比。

③经验法：不同被试团体在测验上的得分和对每题的反应存在较大差异（如一般认为高年级比低年级水平高，若总分随年级增高而增高，则说明有内容效度）。

（4）注意

与表面效度有区分，表面效度是外行人认为某个测验从表面上看好像是测某种心理特质的一种现象，它虽然可以取得被试的合作，但是像人格测验这种为了引出被试真实反应的测验，并不期望有太高的表面效度。

2.结构效度（构想效度、构念效度）

（1）含义：一个测验实际测到所要测量的理论结构或特质的程度，也就是说能够说明心理学理论的某种结构或特质的程度。

（2）用途：适合智力测验、人格测验。

（3）确定方法

一般步骤：提出理论假设、推演有关测验成绩的假设、用逻辑的和实证的方法来验证假设。具体包括：

①测验内部寻找证据：考察内容效度、分析被试答题过程、计算同质性信度。

②测验之间寻找证据

A.相容效度法：求新编测验与某个已知的能有效测量相同特质的旧测验之间的相关，若相关高，则具有较高的结构效度。

B.区分效度法：求新编测验与某个已知的能有效测量不同特质的旧测验之间的相关，若相关高，则说明结构效度不高。

③实证效度法：根据效标将人分为两类，考察其得分差异；根据得分分为高低组，考察其效标差异。若差异显著，则说明有较高的结构效度。

④多种特质-多种方法矩阵法：是相容效度和区分效度法的综合运用。

不同测验测量同一种特质所得相关系数很高，则说明相容效度很高；相似测验测量不同特质所得相关系数很低，则说明区分效度高；相似测验测量相似特质所得相关系数很高，则说明信度较高。

⑤因素分析法包括验证性因素分析和探索性因素分析、结构方程建模及认知心理学上的证据等。

（4）特点

结构效度大小首先取决于心理特质理论；

实测资料无法证实理论时，结构效度并不一定不高，可能是理论假设不成立；

结构效度是通过测量的内容选择积累起来以确定的。

3.实证效度（效标关联效度）

（1）含义：一个测验对处于特定情境中的个体的行为进行估计的有效性。说明应该以实践的效果来作为检验标准（考研选拔的人才，具备高的科研能力，说明实证效度高）。

根据效标资料搜集的时间差异，可分为同时效度（效标资料与测验分数可以同时搜集）和预测效度（效标资料在测验之后根据实际工作成绩来确定）。

（2）用途：同时效度主要用于诊断现状，预测效度主要用于预测某个个体将来的行为。

（3）确定方法

①相关法：测验分数与效标测量的相关。具体相关方法的选取详见心理统计学部分。

②区分法：根据效标测量的好坏分组，回头分析测验分数的差异。

③预期表法：将预测源分数和效标分数制成双维图表，并将每个变量按水平分成若干档次，然后列出每个档次上的人数百分比并从表中看出效标效度的高低。

④命中率及基础率、灵敏度、确认度。

		测验分数（实际测得）	
		高	低
校标（临床表现）	高	正确接受（A）	错误拒绝（C）
	低	错误接受（B）	正确拒绝（D）

假设测验是为了挑选出能力高的人：

正命中率，正确测出的比例：A/（A+B）

负命中率，正确排除的比例：D/（C+D）

总命中率，正确测出和正确排除的比例：（A+D）/（A+B+C+D）

基础率，符合筛选标准的群体占总体的比例：　（A+C）/（A+B+C+D）

灵敏度，符合筛选标准的群体能被筛选出来的比例：A/（A+C）

确认度，不符合要求的群体能被排除的比例：D/（B+D）

当测验为了维护社会公平时，要注重总命中率；当测验用于提高学习或工作效率时，要

注重正命中率。

基础率较低时，应选用灵敏度较高的工具；基础率较高时，应选用确认度较高的工具，这样才能够更有效。比如，当基础率较低时，能力高的人很少，所以要尽量选用灵敏度高的测验将其（高能）筛选出来；而基础率较高时，能力高的人很多，能力低的人就很少了，所以要尽量选用确认度较高的工具将这些人（低能）剔除。

⑤功利率法：一般来说，使用测验所带来的好处应该大于使用测验所耗费的时间、精力和经费。

（4） 效标：效标是衡量一个测验是否有效的外在标准，通常指我们要预测的行为。

常用效标：学业成就、等级评定、临床诊断、专门的训练成绩、实际的工作表现、对团体的区分能力以及其他现成的有效测验。效标可以是连续变量、离散变量、现成指标、人为设计的指标、主观判断、客观判断、主观评价、客观评价，但不能是描述性的资料。

阿斯汀（Astin）将效标分为观念效标和效标测量。观念效标是一个概念，效标测量是观念效标的数量化。

4.总结

内容效度是最适合测量具体属性的测验，如成就测验；

结构效度是最适合测量抽象概念的测验，如自我效能感、人格类型等；

实证效度是最适合用来预测结果的测验，如人事选拔。

（三）影响因素及改进方法

1.影响因素

（1）主试方面

不遵守指导语，评分、计分出错会降低效度。

（2）被试方面

个体的身心状态；团体的同质性，团体需要有必要的同质性。

（3）测量工具方面

样本对预测内容和结构缺乏代表性、指导语不明、题目语义不清、难度过难或过易都会降低效度。另外，测验的长度也会影响效度，关系如下，其中其中 r_{xx} 为原测验的信度系数，r_{xy} 为原测验的效度系数，$r_{(nx)y}$ 为长度相当于原测验 n 倍的效度。

（4）施测过程方面：出现意外干扰。

（5）测验的信度以及效标的选取：信度不高的测验不可能具有很高的效度；测量行为与所选效标相似性越高，效度越高；另外，还需考虑测验分数和效标之间是否有线性关系等问题，以选择正确的相关关系的计算公式。

2.改进方法

（1）主试方面：主试严格执行施测规程，评分者要严格按标准给分。

（2）被试方面：创设标准的应试情境，让每个被试都能发挥正常的水平；选取具有一定同质性的团体。

（3）测量工具方面：精心编制测验量表，避免出现较大的系统误差。

（4）施测过程方面：妥善组织测验，控制随机误差。

（5）其他方面：保证测验的信度，选好正确的效标，定好恰当的效标测量，正确地使

用有关公式。

（四）作用

当测验分数和效标分数呈线性关系时，可运用线性回归的知识，通过测验分数对效标分数进行预测（详见心理统计学部分）：

$$a=\bar{Y}-b\bar{X},\ b=\frac{\sum(X-\bar{X})(Y-\bar{Y})}{\sum(X-\bar{X})^2} \qquad b_{YX}=r\times\frac{S_Y}{S_X},\ b_{XY}=r\times\frac{S_X}{S_Y}$$

$$SE=S_Y\sqrt{1-r_{XY}^2}$$

（五）信度和效度的关系

1.信度高是效度高的必要而非充分的条件。测验效度高，其信度也必然高；但测验信度高，其效度不一定高；但信度低，效度肯定不高。

2.测验的效度受它的信度制约，信度系数的平方根是效度系数的最高限度。

$$r_{xy}\leq\sqrt{r_{xx}}$$

四、心理测验的项目分析

项目分析是根据试测结果对组成测验的各个题目（项目）进行的分析，从而评价题目好坏，对题目进行筛选。

（一）难度

1.含义

难度是指测验项目的难易程度，一般用通过率（P）来表示。一个测验项目，如果大部分人能答对，那么其通过率就高，项目就越容易；一个测验项目，若大部分人都不能答对，则其通过率就低，项目就越难。把 P 理解为“容易度”，更容易做对题。

2.计算

（1）在二分法计分项目中

①通过率：通过一项目的人数百分比。

$$P=R/N$$

其中，P 为项目难度，N 为全体被试数，R 为答对该项目的人数。

P 值越大，题目越容易。

②极端分组法：当被试的人数较多时，可以先将被试分为三组，取最高的 27%被试和最低的 27%被试作为高分组和低分组，并分别计算通过率，最后求两个通过率的平均值作为该项目的难度。其中，P_H 和 P_L 分别表示高分组和低分组的通过率，R_H 和 R_L 分别表示高低组的答对人数，N_H 和 N_L 分别表示高低组的人数。

$$P=\frac{P_H+P_L}{2}=\frac{R_H/N_H+R_L/N_L}{2}$$

（2）在非二分法计分项目中

$$P=\bar{x}/x_{max}$$

其中，$\bar{x}$为所有被试在该项目上的平均得分，x_{max} 为该项目的满分。

（3）选择题的猜测校正

进行猜测校正是为了排除由于猜测而答对某些题目致使通过率增大的可能性。包括：

①全体被试在某个项目上的通过率的校正

$$CP=KP-1/K-1$$

其中，CP 为校正后的难度，P 为校正前的难度，K 为选项的数目。

②某个被试参加多个项目组成的测验的测验分数的校正

$$S=R-W/K-1$$

其中，S 为校正后的得分，R 为答对的项目，W 为答错的项目，K 为选项的数目。

3.难度水平的确定

项目难度水平的确定取决于测验的目的和性质：

（1）对于常模参照测验，项目难度应尽量接近 0. 50，以尽可能地区分被试个体差异；

（2）对于标准参照测验和掌握测验，可不考虑难度；

（3）对于选拔和录用测验，应将测验的项目难度控制在录取率左右；

（4）对于选择题，难度应该大于猜测概率；

（5）速度测验难度不宜太高且各难度应接近，难度测验要求难度在 0. 50 左右；

（6）人格、态度、心理健康等测验一般不需要难度。

总之，无论何种测验，一般都应防止被试得满分，因为满分的意义是不明确的。

4.难度对测验的影响

（1）影响测验分数的分布形态

①难度较大的测验，通过率低，测验分数集中在低分端，只有少数人得高分，分数分布将呈现为正偏态，正偏态分布适合于筛选性测验（如：英语竞赛）。

②难度较小的测验，通过率高，测验分数集中在高分端，只有少数人得低分，分数分布将呈现为负偏态，负偏态分布适合于达标性测验（如：中学会考）。

③中等难度的测验（被试取样有代表性的情况下），分数分布呈现正态分布。

一般能力测验和成就测验的平均难度在 0.50 左右为宜。

（2）影响测验分数的离散程度和信度

过难或过易的测验会使测验分数相对地集中在低分端或高分端，从而使分数的全距缩小，信度降低。

项目的难度以集中在 0.50 左右为最佳，以集中在两端为最差。

5.难度的等距转换

难度用通过率来表示计算方便，便于理解，但它是属于顺序变量，没有相等单位，这为进一步分析带来困难；另外，难度量表是反序而行，P 值越大，项目越容易。因此需要设法将其转换为等距变量。

可以将难度转换为 Z 分数，为了克服 Z 分数有小数点和负值的缺点，也可转换为另外的难度指标：$\Delta=13+4Z$ 或 $Z'=Z+5$。

（二）区分度

1.含义

区分度是指测验项目对被试心理品质水平差异的区分能力或鉴别能力，被用作评价项目

质量，是筛选项目的主要指标和依据。

区分度的高低依赖于被试水平的准确测量，一般称为效标分数，但效标分数更多使用测验总分，又称内部效标。

通常用 D 表示，D 取值在[-1.00，+1.00]之间，D 为正值是积极区分，且 D 越大区分效果越好；D 为负值是消极区分；D 为 0 是无区分作用。

2.计算

（1）项目鉴别指数法

当效标成绩是连续变量时，可以从分数的两端各选择 27%的被试，分别计算出每道题目上各自的通过率，两者之差便是鉴别指数（D）。D 值越高，区分度越高，即项目越有效。

$$D=P_H-P_L$$

美国测量学家伊贝尔（L. Ebel）提出：D>0.40，题目很好；D 值在 0.30~0.39 之间，题目良好，修改会更好；D 值在 0. 20~0.29 之间，题目尚可，仍需修改；D<0.19，题目差，必须淘汰。

（2）相关法

以项目分数与效标分数或测验总分的相关作为项目区分度的指标。项目数量与测验总分之间，相关越高，项目的区分度越高。根据不同的情况，可以使用点二列相关、二列相关、φ相关和积差相关等。具体方法见心理统计学部分。

（3）方差法

被试在某一项目上的得分越分散，即方差越大，则该试题鉴别潜力越大。

3.区分度的相对性

（1）不同计算方法，所得区分度不同。一个测验的各个项目要采用同一种区分度指标。

（2）样本容量大小影响相关法计算区分度值的大小。一般来说，样本容量越小，其统计值越不可靠。

（3）分组标准影响鉴别指数。分组越极端，其 D 值越大。

（4）被试样本的同质性程度影响区分度值的大小。被试团体越同质，即个体之间水平越接近，其测验题目的区分度值越小。

4.区分度与难度的关系

（1）难度越接近 0.5 时，项目潜在的区分度越大；

（2）难度越接近 1. 00 或 0 时，项目区分度越小；

（3）为了使项目具有较高的区分能力，应该使所有项目难度都保持在 0.5，但是从整体来说，这样做会使测验所提供的信息相对减少。所以，应使项目的难度分布广一些，梯度大一些，使整个测验的难度分布呈正态分布，且平均水平保持在 0.5 左右。这样才能把各种水平的人都区分开来，并且分得比较细。

（三）题目的综合分析和筛选

1.看区分度

低区分度的题目是不能有效鉴别被试的，一般来说，0.3 以上是比较好的。另外，考虑到区分度的相对性、评价项目的有效性时，应考虑到测验的目的、功能以及被试团体的总体水平，不能将区分度作为筛选试题的绝对标准。

2.看难度

难度一般在 0.35 到 0.65 之间较好，但就整个测验而言，难度为 0.5 的题目应居多，也需要一些难度较大和较小的测题，使难度成一个均值为 0.5 的正态分布。

如果是人格测验、态度测验以及心理健康测验等，难度一般较低，以保证每个被试能理解测题的意思；如果是标准参照测验，则应该根据编制测验时确定的目标来选择难度。

根据区分度和难度水平选择出合适的测题后，应该与根据原来双向细目表所选的测题相对照，看它们之间的比例是否失调，如果失调，应加以调整。

3.选项分析

对选择题后面提供的几个答案的分析。此时主要的异常情况有：

正确答案无人选择，或少于其他选项的人数；错误答案选的人太多；正确选项上的高分组选择人数少于低分组；错误选项上的高分组选择人数又多于低分组；某个选项无人选择；未答的人数较多等。

4.分析原因，酌情修改

不要轻易丢弃不符合要求的项目，因为:

（1）用内部一致性分析所求得的区分度不一定能代表试题的效度。

（2）区分度指数低的试题不一定表示该题有缺点。要详细分析区分度低的原因，并保留题目，作为测验一项重要的学习结果的记录，以备日后使用。

（3）课堂测验的项目分析资料的有效性是随时空而变化的，并非固定不变的。

（4）编制新的项目需要的时间几乎比修订现存项目长 5 倍。另外，如果做因素分析，还要看题目的负荷量与题目间的相关，某个因素中题目过少的，也要进行删除；题目的筛选也要考虑量表的长度，一个测验的长度应该根据测验的时限、对象的年龄、测验的性质而定。

五、经典测量理论的评价

（一）经典测量理论的优点

1.CTT 以随机抽样理论为基础，建立在简单的数学模型之上，直观易懂，易于被理解和接受，计算也简便，容易推广。

2.理论假设较弱，对实施条件要求不严格，适用性广，对许多测验结果都方便分析。

3.多数情况下，CTT 还是足够精确的，测验结果是可信的。

（二）经典测量理论的局限

1.对信度的估计精度不高，平行测验难以实现。2.误差指标笼统单一，难以精确计算个体被试的独立误差。3.各种参数的估计对样本抽样的依赖性太大，而获得代表性样本却很困难。4.参数指标之间的配套性较差，与被试水平参数之间关系模糊（参数与被试水平不在同一参照系上）。5.真分数与观察分数之间存在线性关系的假定不合理。6.不太适合标准参照测验。