

扩展知识：

### 一．控制题目难度的基本方法

一般来说影响题目难度的主要因素有

- 1.考查知识点的多少；
- 2.考查能力的复杂程度或层次的高低；
- 3.考查对题目的熟悉程度（如本来较易的题目会因考生均为之一而造成很难，或本来较难的题目会因为考生普遍练习变的较容易）；
- 4.命题的技巧性（如同一个问题，可以，命得容易，也可以较难）

控制题目因素除了考虑上述因素，还可以通过其他方法来控制。在平常的教学考试中，由于老师对学生的情况比较了解，因而只要凭经验来控制难度，使之与老师难度相适应。而在大规模的测试中，就要通过预测来掌握难度了，首先由命题人员根据上述因素估计一个难度范围，然后通过测试看这个估计的准确程度分析原因，进而提高评估能力。经过预测取得难度的题目可以进入题库，以后备用。

（摘自《教育测量与评价》胡中锋、李方编著，广东高等教育出版，p54）

### 二．黄金法则案例

在 20 世纪 80 年代早期，黄金法则保险公司控告了伊利诺伊保险部门以及教育考试中心（Educational Testing Service, EST）黄金法则的诉讼指出，EST 为伊利诺伊保险部门开发的两个保险执照测验对于黑人被试存在偏差。他们所说的偏差指的是，白人在很多项目上的正确作答率比黑人的都高。该诉讼所依据的统计量是 P 值，即用于表明项目难度的统计量。这种对于黑人和白人不同的项目难度水平，导致了大量的白人而不是黑人通过该测验并且拿到了执照。该案例最终庭外和解，双方达成协议，EST 将开发出进一步的测验，在新测验中的项目，所有被试的正确作答概率都低于 0.40。而且，双方协定，如果黑人和白人在某个项目上的正确作答率相差超过 15%，那么该项目将不会被用于测验。

很多测验专家提出，黄金法则案例的解决方法是有问题的。他们指出，通过 P 值对项目难度进行比较的方法，并没有将被试的能力、水平考虑近来。换句话说，被试没有正确作答是因为项目出的不好，还是仅仅因为被试不知道如何准确回答？

正如研究者们在美国心理学会（APA）年会的一个讨论上，以及 APA 的一个官方声明中所指出的，项目反应理论(IRT)提供了更好的方法以识别项目是否对一个群体比另一个群体更困难。你或许还记得，ICCs 在项目难度与被试的能力之间建立起了联系。对于 ICC 进行比较

的是项目难度更好的方式，因为它将被试的能力考虑进来。而且，Rodney Lim 在他的硕士论文中，对这个原则进行了实证的实验。他使用这两种方法对同一组模拟数据进行了评价。他发现黄金法则的结果是 90% 的项目是存在偏差的，相反，IRT 的方法在发现，当我们将能力因素考虑进来的时候，只有 25% 的项目有偏差。

### 三.文化适应测评的作用

当我们进行一个定量的项目分析时，我们的目的是识别那些因此测验本身、评分设计、施测或解释等原因而导致最终测验成绩出现误差的来源。因此，我们对被试在测验上的反应进行考察以确定是否存在某个项目对某个群体而言更容易的现象，并且，我们通过对被试在测验上的反应进行考察，以确保每个项目对被试的区分仅仅是以待测构想为基础的。

直到最近，那些对项目进行文化偏差分析的测验开发者通过将少数民族的被试包括进常模样本中，并且通过使用项目特征曲线（ICC）或者定性的方法识别偏差项目。在这个阶段，一旦发现某个项目存在文化偏差，那么该项目或者被修改或者被删除。尽管这个方法很好，但是，在一个变化多样的群体中要想识别文化偏差，这么做是不充分的。21 世纪的美国和欧洲已经具备了真正的多元文化的特征，很多个体的行为和传统拥有两个甚至更多的文化烙印。而且一个人或许对于两个或多个文化的适应程度也是不一样的。例如，一个年轻人生长在二元文化家庭中，墨西哥-美国或中国-美国，他对其中一种文化或许更认同；相反，另一个生长于同样的二元文化家庭中的年轻人，或许对另外一种文化更认同。在多元文化的人群中，少数民族群体内部的成员各自的文化适应程度也不尽相同。在这样的背景下，对每一个少数民族群体进行项目分析往往让人觉得非常棘手，而且很有可能是无效的。

来自荷兰的两个研究者，采取一种不同的方法解决这种与被试对主流文化的理解和遵从程度有关的项目或测验偏差，他们建议首先对每个被试对主流文化的融入程度进行测评，然后使用这些测验分数对原始分进行校正，从而剔除测验分数中由于适应不足而导致的误差。现在对适应程度进行测量的方法多为调查量表（大约 20 个项目），对被试感觉自己归属一个或多个群体的程度，以及他们对待这些群体的态度进行考察。

根据 Vvan de Vijver 和 Phalet(2004)的研究，测验中存在三种类型的文化偏差：

1. 构想偏差（constuct）：当同一个项目在不同的文化或亚文化中具有不同的含义时，就会出现这种偏差。例如，在不同的文化中，做一个好女儿和好儿子的行为就存在很大的不同。
2. 方法偏差（method bias）：当测验的技术在不同的文化群体中存在不同时，会出现这类偏差。例如，在北美和欧洲上学的人往往对使用电子扫描答题卡或计算机文化更为熟悉，而那些习惯书面或口头的方式回答施测者的问题的人则不习惯于那些方法。

3.项目功能差异（**differential item functioning**）:当来自不同文化背景但具有相同能力水平的被试，因为文化差异而在项目上火测验得到不同的分数时，就是出现了偏差。例如，一个历史测验包含了大量有关美国最早的 13 个殖民地的历史和地理问题。这样的测验对于那些来自新英格兰的被试则更有利，而那些来自加利福尼亚或墨西哥的被试则会觉得困难，尽管这两个群体对于美国整体的历史同样熟悉。