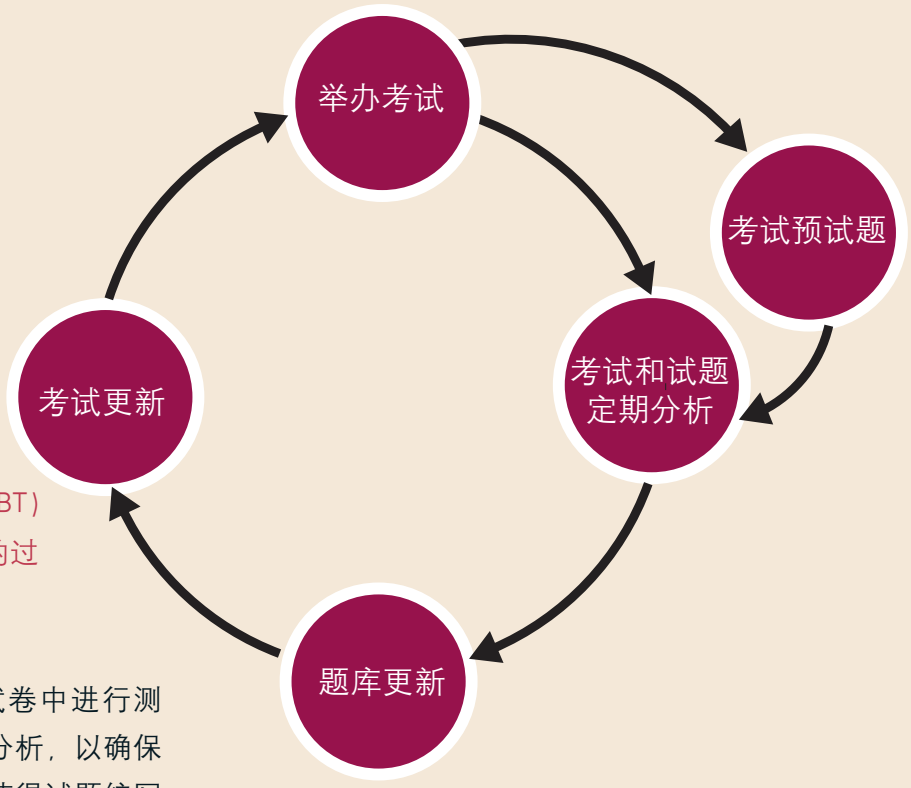
评估和保持考试质量



对于可随时约考的计算机化考试（CBT）来说，试题和考试开发是一个连续循环的过程，如右图所示。

预试题作为不计分题被放入正式试卷中进行测试。应定期对试题和考试数据进行全面分析，以确保考试保持应有的质量标准。这些分析也使得试题编写人员了解试题表现，帮助其做出正确的组卷决策。通过以上方式不断对题库进行试题补充，并定期更新考试。

心理测量分析

为确保试题达到应有的质量标准，Pearson VUE的心理测量师不断对考试和试题的表现进行监测。这包括定期监测及格/不及格率，校验试题统计参数漂移，了解试题是否会随着时间推移变得更容易或更难。需要监测的各类考试和试题数据包括：

评分的标准差
- 应试者数量
- 正确作答的应试者数量
- 试题难度
- 标准差
- 试题区分度
- 每道试题的答题时间

选项（干扰选项分析）
- 应试者数量
- 正确作答的应试者数量
- 选项难度
- 选项区分度

考试/题库
- 按内容的试题比例
- 按难度的试题比例
- 按内容的难度比例

我们的心理测量师定期提交技术报告，使考试主办方及各利益相关方了解考试和试题的表现情况。该技术报告通常含有试题级别的信息、通过分数线信息、信度、原始和量表分数以及问卷等值化程序等信息。本文后续的章节将就技术报告中的一些内容做深入的讨论。

信度和评分标准差

考试总会有评分误差，这并不意味着在考试制作或评分的过程中出现了错误，而是考生参加不同场次的同一考试且使用同一问卷时成绩并不完全一致，此类误差也存在于不同的问卷之间。我们用信度系数来估计评分误差的程度。信度系数和评分标准差是衡量考试分数准确性的两个最常用的指标。

信度系数有多种类型，分别用来评估不同类型的评分误差。其中一个最常用的系数是内部一致性信度系数。该系数根据不同考生在某一次考试中对同一问卷的作答情况来评估误差。其它类型的信度系数，则根据考生在不同考试中或作答不同考卷时的表现来评估误差。

信度系数的取值范围介于0.00到1.0之间，通常在.70到.95的范围内。信度系数越接近最大值，考试评分误差就越小。影响内部一致性信度系数的考试特征有试卷长度（试题数量）、答题时间和试题之间的关联性。通常情况下，允许考生有足够时间答完所有考题的费时较长的考试，和那些评估单一方面知识或技能的考试，内部一致性可信水平较高。我们在解读信度系数或在不同考试之间比较该值时，需要考虑到这些因素。

评分标准差（SEM）是由前面所说的评分误差造成的分数与应试者实际或“真实”掌握知识水平之间的差距。为了理解SEM的含义，不妨作一假设：一位数学知识达到一定程度的考生，多次参加同一数学考试。即使在没有记忆试题，且他/她的知识水平也没有变化的前提下，这位应试者也不太可能获得完全一致的分数，而会取得接近其实际知识水平的相近的成绩。这些成绩之间的差异，即测量到的结果和实际水平间的差距，正是SEM所要评估的。具体来讲，SEM就是这些差异分布的标准差。评分标准差和信度成反比关系，即考试信度系数越高，则评分标准差越小，反之亦然。

我们很难确定考生真实水平和其考试成绩之间的差距，但测评方面的理论使我们能够应用SEM来建立应试者分数区间或置信区间。在特定合理的假设下，应试者的考试成绩有68%的概率处于其真实水平的±1标准差之内，有95%的概率处于其真实水平的±2标准差之内。

试题分析

应至少运用经典测试理论（CTT）对试题进行分析和报告。经典测试理论的两大指数为难度和区分度。可以根据试题分析结果标记相关试题，以供命题专家审阅。

名词解释

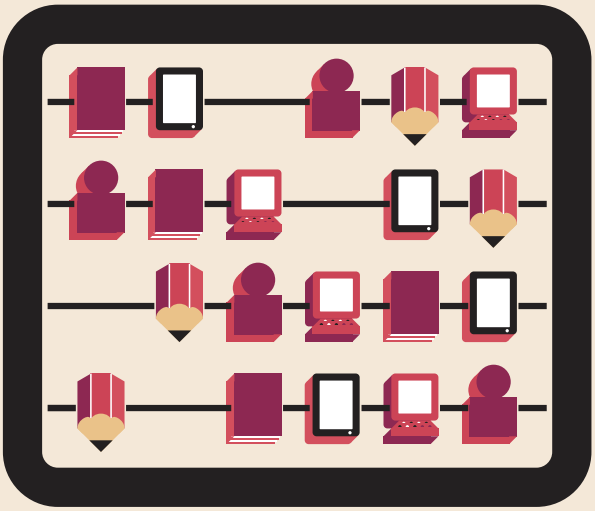
- **量表信度值分析：**
是一个用来评估考试分数可靠性的统计指标。阿尔法测量的是内部一致性，即试题评估同一知识或技能的程度。在特定合理的假设下，阿尔法也用来评估应试者对同一考试不同试卷的表现相似度。

- **干扰选项分析**
是对选择题干扰选项表现好坏的统计分析。

- **难度**
用以测量试题的难易度，在经典考试理论中又称p值。

- **区分度**
答对单个考题和通过整体考试之间的关联性。

- **试题分析**
对应试者的答题情况进行统计分析，以了解试题的质量情况。



经典测试理论中的试题难度仅是正确作答的考生比例，由此得出的统计结果（称为p值）介于0到1.0之间，p值越高，试题难度越低。尽管考试完全可以包含极易或极难的试题，但一般不应出现p值小于0.20或大于0.95的四选项多选题。通常情况下，考试应包含不同p值的试题，不宜有太多的易题或难题。

试题区分度用以衡量单题分数和考试成绩的相关性。这种相关性反映了单题表现和整体考试表现的关联程度。统计结果介于–1.0到1.0之间。较好的统计结果是高度正相关（例如：大于0.20），而负相关则说明那些考试总体表现不太好的考生在该题表现较好。

根据考试目的和考试方案不同，还可以进行其他类型的试题分析，如试题反映理论（IRT）和试题表现差异理论（DIF）。试题反映理论是描述应试者答题表现和能力水平之间关系的数学模型。试题表现差异是“一种试题统计特征，用以研究考试成绩相同的不同组别考生的单题平均分差异，或选择不同选项的比例差异”。（美国教育研究协会（AERA）、美国心理学协会（APA）和美国国家教育测量委员会（1999）教育和心理测试标准，第175页）。

内容审核

内容专家对标明统计特征的试题进行审核，做出诊断性评估并确定组卷方案。下表列出了一些较差试题的统计学特征：

试题统计特征	
低p值	负或低相关性
潜在原因：	
答案不正确 有一个以上的正确答案 试题内容生僻或过于琐细 试题表述不清晰	答案不正确 有一个以上的正确答案 试题太难，应试者在猜答案 试题表述不清晰 试题测试内容与其它试题无关

内容审核后，试题可能被：1）准许继续使用；2）重新编写，并重新测试；3）撤下并在题库中存档。

题库

使用题库能够追踪试题，提供试题表现的历史记录，并可根据试题特征（包括统计特征）进行检索。Pearson VUE内容开发人员对题库的管理包括根据考试蓝本监控考题的内容和难度比例。 同时，通过题库差距分析来指导试题编写工作。

完整的考试维护计划包括统计分析、根据试题表现进行内容审核和有针对性的试题编写工作，这是高质量考试的有效保障。

• 库德·理查森公式 (Kuder–Richardson Formula 20)
与量表信度值分析相同的是，库德·理查森公式也用于评估内部一致性信度。不同的是，KR20只能用于二分测量法。

• 信度
监测不同条件下同一组考生考试成绩的一致性和稳定性，这些条件可能包括不同的考试时间，发送模式，不同的试卷和样题。

• 评分标准差（SEM）
可反映一组考试分数的随机误差的大小，用来评估由于各种随机因素导致的考生分数的变化趋势，如考生所抽中的试卷中的考题。评分标准差越小，这些因素的影响就越小。

参考资料：
美国教育研究协会、美国心理学协会和国家教育测评委员会（1999）；教育和心理测试标准；华盛顿特区：美国教育研究协会。

国际测试委员会（2001）；国际测试应用指南。http://www.intestcom.org/ test_use.htm（2012年1月7日检索）。

Pearson VUE销售办公室

美洲

芝加哥，伊利诺斯州
+01 888 627 7357
pvamericassales@pearson.com
www.pearsonvue.com

全球总部
明尼阿波利斯市，明尼苏达州
+01 888 627 7357
pvamericassales@pearson.com
www.pearsonvue.com

费城，宾夕法尼亚州
+01 610 617 9300
pvamericassales@pearson.com
www.pearsonvue.com

欧洲、中东和非洲

迪拜，阿联酋
+971 44 535300
vuemarketing@pearson.com
www.pearsonvue.ae

伦敦，英国
+44 0 207 775 6737
vuemarketing@pearson.com
www.pearsonvue.co.uk

曼彻斯特，英国
+44 0 161 855 7000
vuemarketing@pearson.com
www.pearsonvue.co.uk

亚太区

北京，中国
+86 10 5989 2600
pvchinasales@pearson.com
www.pearsonvue.com.cn

德里，印度
+91 120 4001600
pvindiasales@pearson.com
www.pearsonvue.com

墨尔本，澳大利亚
+61 3 9811 2400
pvseasiasales@pearson.com
www.pearsonvue.com

东京，日本
+81 3 6891 0500
pvjsales@pearson.com
www.pearsonvue.com/japan

持续评估和保持考试质量

如需了解更多信息，请访问：www.pearsonvue.com