响等值精度。③单组锚人：保障施测安全性。施测成本较高。可根据锚人数量估计抽样误差，保障等值精度。**等值方法**：①同时标定：操作过程简便。对共同题的参数标定精度较高。缺失值影响参数估计结果。②分开标定：对计算机运算要求低。可检验锚题是否存在参数漂移。存在量尺转换误差。③固定参数标定：操作过程简便。锚题数量对等值结果的影响较大。

3. 标准设定。概念：包括内容标准与表现标准。内容标准是学生必须掌握的知识、技能与能力，回答“是什么”的问题。表现标准是反映学生所掌握内容与技能的测验分数，回答“是多少”问题。**作用**：一是对学生进行分类和解释（合格 / 不合格或基础 / 熟练 / 高级）。二是评价考生是否达到进入下一阶段学习的要求或是否具备从业资格。三是有助于形成考试行业标准与规范。**方法**：一是安戈夫法。特点是：最早提出且被广泛使用，基于 CTT。拥有很多变式。操作简单、使用方便。相对于开放题，更适合选择题。二是书签法。特点是：基于 IRT，将分界分数的设定与测量模型建立联系。更适用于混合格式测验。 题目按照 BDL 由小到大排列，可减轻评委的认知任务，在设定分界分数时还可减少评委判断次数。 可方便地在单个测验中设置多个分界分数。

——北京师范大学副教授陈平

学业水平考试的测量学原理与方法

教育测量学的理论和国际考试行业的成功经验已经表明，要做好高中学业水平考试，必须提前至少一年做好四项测量学准备，即研发标杆试卷，确定测量标准，研发测验常模和实施测验等值。

一、标杆试卷的研发。首要是确定考查的内容标准，精心编制每个学科的考试说明。**其次**要编制考试蓝图。考试蓝图类似于工程建设中的施工图纸，施工时的材料（考试题目）和结构（试卷编排）等必须完全符合图纸要求，这是为日后的测验等值打基础。**第三**，制定组卷方案。**第四**，开展题目质量评价。**第五**，开展组卷质量评价。**第六**，开展信度效度检查。

二、测量标准的确定。包括内容标准的设定和表现标准的设定。其中，内容标准的设定在研发标杆试卷时已经完成。表现标准的设定需要有 10 个步骤：**第一步**，组建专家小组。**第二步**，选定标准方法。**第三步**，编写等级说明。**第四步**，学习设定方法。**第五步**，初步评定等级。**第六步**，综合多方信息。**第七步**，反复多轮讨论。**第八步**，终审评定结果。**第九步**，评估评定过程。**第十步**，收集效度证据。

三、测验常模的研发。确保实现四个目标。**第一**，样本代表性好。要充分考虑人口数据、历史数据、变量选择等。**第二**，确保信度较高。测量的随机误差需要得到有效控制。**第三**，参照点和单位要合理。要合理选择参照领域分数、年级 / 年龄参照分数、成长分数等。**第四**，常模曲线合理且拟合良好。

四、测验等值的实施。新高考试行一年多考是一项具有划时代意义的进步举措，但前提是能够实现测验分数的等值。可尝试事后等值设计方法。基本思路是：**首先**根据标杆试卷 R，编制一个锚测验 A。**其次**，选择一个能力跨度与全省能力跨度十分接近的全省考生的代表性小样本（530~1600 人），让该样本中的一半考生按照 R−A 方式作答，另一半考生按照 A−R 方式作答，于是可以获得一个 RA 的考生作答反应数据。**第三**，待全省学业水平考试 F 正式施测之后，从中选出另一个全省考生的代表性小样本（530~1600 人），让他们在学业水平考试之后 2 天之内单独考一次锚测验 A，于是可得另一个数据 FA。**第四**，综合数据 RA 和 FA，利用锚测验题目参数不变性特点，采用项目反应理论或经典测验理论等方法，实现正式测验 F 与标杆试卷 R 之间的测验等值，即建立正式测验的原始分数与标杆试卷原始分数之间的对应关系。

——湖南师范大学测评研究中心主任杨志明

江苏教育考试科研月报

2020 年第 12 期（总第 72 期）

江苏省教育考试院编印

“教育测量与评价”研修专题

让教育测量回归正确的评价导向

——我省举办“教育测量与评价”专题线上研修班

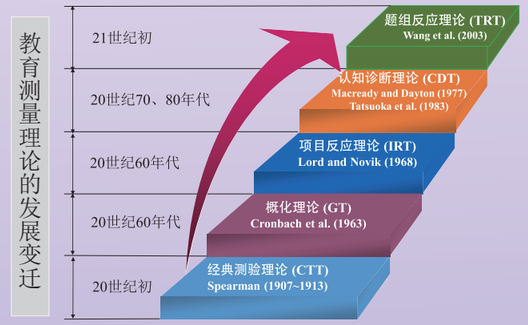
改进结果评价、强化过程评价、探索增值评价、健全综合评价，着力破除唯分数、唯升学、唯文凭、唯论文、唯帽子的顽瘴痼疾，建立科学的、符合时代要求的教育评价制度和机制，这是我国新一轮教育评价改革发展的总目标。要实现这一目标，不仅需要从政策和管理层面作出周密和系统的安排，而且需要从教育测量与评价的理论与技术层面拿出切实可行的实操方案。2020 年 10 月 29 日至 11 月 20 日，我院与清华大学继续教育学院联合举办了为期四周的“教育测量与评价”专题线上研修班，为全面提升全省招考系统测量与评价专业能力搭建平台，为全面落实国家教育评价改革总要求奠定基础。来自全省各市招考系统干部 100 余人参加了培训。

教育评价对教育实践有着重要的影响，起着指挥棒的作用。因此，教育评价改革显得尤为迫切，是建设现代教育治理体系、提升教育治理能力的关键。中国教育在线总编辑陈志文从高考改革角度切入，分析分享了教育评价改革面临的挑战与思考。湖南师范大学测评研究中心主任杨志明阐述了要做好高中学业水平考试，必须要做好标杆试卷的研发、测量标准的确定、分数量表的设计以及测验常模的研发四项测量学准备。北京师范大学中国基础教育治理检测协同创新中心副教授陈平介绍了教育测量的三种基本理论及三项关键技术，并举例说明教育测量的数据应用。除了教育测量领域的专业分享，上海师范大学原校长张民选从专业角度全面介绍了国际学生评估项目（PISA）；清华大学招生办主任余潇潇介绍了清华大学落实国家强基计划的探索与实践；清华大学书院管理中心主任苏芃介绍了清华大学推进本科教育改革、提高人才培养质量的积极探索；教育部职成司高等教育处处长邬跃分析了继续教育的政策与形势。这些学术分享，从不同角度为教育评价改革发展提供了借鉴与思考。

教育测量是教育评价的前提和基础， 如果没有教育测量的事实判断，教育评价就成为无水之源，失去了价值判断的基本依据。反之，教育测量的结果也只有通过教育评价这个环节才能获得实际的意义。让教育测量回归正确的评价导向，应该是教育测量的旨归所在。

【编者按】中共中央、国务院印发的《深化新时代教育评价改革总体方案》提出要“改进结果评价、强化过程评价、探索增值评价、健全综合评价”，阐明了新时代教育评价改革具体方略。其中，“改进结果评价、强化过程评价”反映教育评价从结果导向到过程导向的转变。政策层面教育评价改革思潮的变化与教育评价理论与测量方法技术发展趋向息息相关。本期，依托与清华大学开展的“教育测量与评价”线上培训资料库，为读者介绍教育测量理论与技术相关内容。

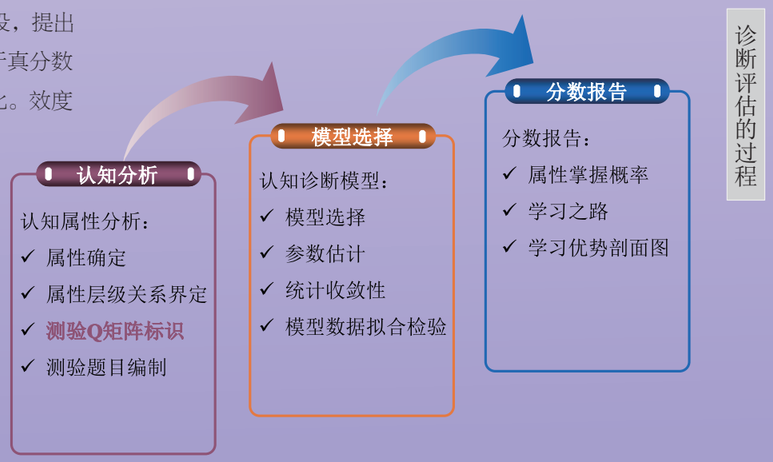
教育测量的几种基本理论介绍



1. 经典测验理论（CTT）。亦称“真分数理论”。20 世纪初提出，至 50 年代臻于完善。该理论假设观察分数线性表示为真分数与误差分数的和，即 X= T+E。假设某个人的潜在特质可用多个平行测验反复测量多次，则所得观察分数的平均值会等于真分数，即 E(X) =T。假设真分数与误差分数不相关，即 ρ (T，E)=0。不同平行测验上的误差分数不相关，即 ρ (Ei，Ej) =0 (i ≠ j)。根据这些基本假设，提出信度和效度的概念。信度等于真分数变异数与实得分数变异数之比。效度等于有效分数变异数与实得分数变异数之比。**局限性**在于：一是对于参加同一测验的被试，其分数的测量标准误相同。二是只有平行测验的分数才能进行合理比较。三是参数的无偏估计依赖于代表性的样本。

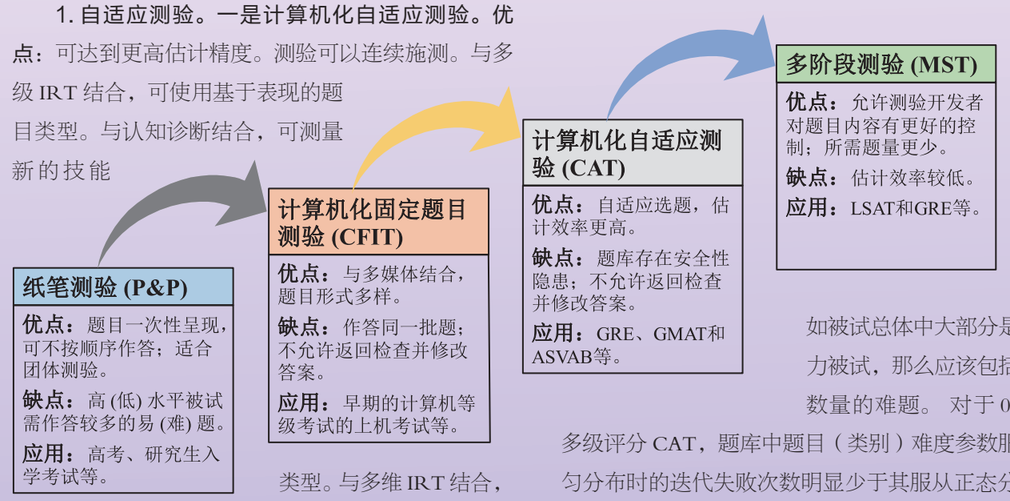
2. 项目反应理论（IRT）。是新兴的教育测量理论，是一种现代心理测量理论，是在批评 CTT 局限性的基础上发展起来，由美国测量学家 Lord 于 1952 年首次提出。它一般适用于难度测验而非速度测验。IRT 将可观察的被试行为（作答反应）与不可观察的潜在特质（ θ ）联系起来，并将这种关系模型化和参数化。相较于 CTT，有以下**优点**：一是不同能力被试有不同的测量标准误。二是被试间的测验难度水平不同也可进行分数的直接比较。三是非代表性样本也能获得参数的无偏估计。

3. 认知诊断理论（CDT）。对学生在测验所测属性 (比如通分、借位与约分等) 上的掌握水平进行分类。通过认知诊断确定学生的认知结构或知识状态：学生掌握哪些属性，哪些属性未掌握需要补救。**诊断过程：**认知属性分析→模型选择→分数报告。

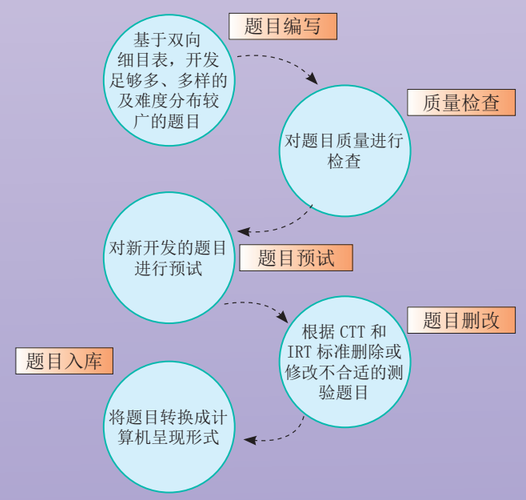


教育测量的几项关键技术介绍

测验形式的发展变迁



1. 自适应测验。一是计算机化自适应测验。优点：可达到更高估计精度。测验可以连续施测。与多级 IRT 结合，可使用基于表现的题目类型。与认知诊断结合，可测量新的技能类型。与多维 IRT 结合，可提供被试在各分维度上的精细信息。任意两名被试在同一时间不可能作答相同题目，抄袭收益不大。**缺点：**为获得稳定的题目特征，所有题目预先施测于较大的代表性样本。题库管理与维护充满挑战。 大多数 CAT 不允许被试检查并修改答案。 CAT 提供连续测验，题库存在安全性隐患 (大规模网上偷题)。 CAT 与 P&P 的兼容性问题：CAT 得分可能低于 P&P 得分。 测验开发者无法提前“过目”题目。 高风险测验中，被试作答不同题目集，可能涉及公平问题。**二是多阶段测验。优点：**允许测验开发者对测验内容有更好的控制，如内容领域、认知技能、题目类型等方面的均衡；题库所需题量更少。**缺点：**测验长度相同的前提下，MST 的估计效率低于 CAT。 **三是题库建设。题库大小：**对于 0−1 评分 CAT，建议题库大小超过 100 题；对于高风险或要求内容均衡的 CAT，建议题库大小在 500−1000 (Dodd, De Ayala, & Koch, 1995)。对于多级评分 CAT，题库大小在 30 题左右就可获得较准确的能力估计值且迭代不收敛情况较少 (Dodd, Koch, & De Ayala, 1989)。Stocking (1994) 建议题库大小至少应该是测验长度的 12 倍左右，Way (1998) 称之为一种经验法则。**题库结构:** 应合理覆盖整个能力和内容范围。依赖于被试总体。如被试总体中大部分是高能力被试，那么应该包括较大数量的难题。 对于 0−1 或多级评分 CAT，题库中题目（类别）难度参数服从均匀分布时的迭代失败次数明显少于其服从正态分布时的结果。**题库开发:** 题库并不是对大量题目的简单堆积，而需要依托专业的团队使用科学的方法进行构建 。



2. 跨年等值。研究目标：实现相邻监测周期相同学科的分数可比。实现“监测周期大于 2”的跨年度等值。得到各学科学生能力的发展变化趋势。**基本原则：**可操作性原则、精准性原则、安全性原则。**等值设计：**①锚题设计：节约施测成本。等值精确度较高。等值结果受锚题质量影响较大。影响测试安全性。②等组锚人：保障测试安全性。施测成本最高。抽样误差影