江苏教育考试科研月报

2024年第9、10期（总第116、117期）　　江苏省教育考试院编印

本期内容简介

【理论概览】　系统介绍四种现代教育测量基本理论和方法：经典测量理论（CTT）、项目反应理论（IRT）、概化理论（GT）、认知诊断理论（CDT）。

【实践应用】　以高中学业水平等级性考试、研究生招生考试、中小学教师资格考试以及全国英语等级考试等作为实例，阐述相关教育理论在教育考试领域中的具体应用。

【成果展示】　介绍我院在教育测量理论应用于优化考试结果评价方面的研究成果。

教育测量理论的发展与应用

编者按：测量理论是研究如何通过测量工具对测量对象的心理特质进行量化描述的理论体系。新一轮考试招生改革的深入推进，对教育考试的测量与评价提出了更加科学化的要求。怎样服务国家战略需要、评价选拔各类人才、维护教育考试评价公平公正，是教育考试测量与评价领域的重大使命。本期月报聚焦现代教育测量基本理论和方法，通过介绍经典测量理论、项目反应理论、概化理论、认知诊断理论在教育考试领域中的应用，展现了教育测量在完善考试结果评价方面的重要影响。

【理论概览】

现代教育测量基本理论和方法

测量理论是研究如何通过测量工具对测量对象的心理特质进行量化描述的理论体系。其核心在于将观察到的现象通过数据的方式进行描述，即对事物作出量化描述。就其内容来说，包括测量对象的确定、测量规则选用、测量工具制作、测量结果分析。心理和教育测量从19世纪末诞生到现在，其理论不断发展和丰富。20世纪50年代之前，经典测量理论（Classical Test Theory，CTT）在相当长的一段时期里稳步发展且广泛应用，成为指导教育测量的主流方法和核心理论；50年代后，项目反应理论（Item Response Theory，IRT）、概化理论（Generalizability

Theory，GT）、认知诊断理论（Cognitively Diagnostic Theory，CDT）逐渐发展。

一、经典测量理论（CTT）

CTT 又称真分数理论，基本假设为所有测量都存在误差，观察分数与真分数之间是一种线性关系，且只相差一个随机误差，其数学模型为：$X=T+E$。其中 X 是观察分数，T 是真分数，E 是随机误差。**CTT 认为，只要平行测验次数足够多，观察分数的平均值会无限接近真分数**。当前，教育考试中常用的**信度、效度、难度、区分度等“四度”指标大多基于经典测量理论得出**。

CTT 是目前测验学界使用与流通最广的理论，优点是理论体系成熟，分析方法简单，意义直观明了，易于理解和掌握，但同时存在着难以克服的技术问题：①测验性能指标估计严重依赖被试样本；②测验总分由各试题的观察分数等权重直接累加；③测量误差估计的不精确性和笼统性；④被试能力与项目难度两个指标的不统一性。这些缺点与 CTT 基于弱假设有关，在其理论体系内部很难得到解决。

二、项目反应理论（IRT）

IRT 又称潜在特质理论，它不再以整个测验为考察对象，而以项目（试题）为考察对象，并在此基础上去分析作为测验基础的项目与由它所构成的整个测验的关系。项目反应理论建立在 2 个基本概念上：（1）考生在某一测试试题上的表现情形，可由一组因素来加以预测或解释，这组因素叫做潜在特质或能力；（2）考生的表现情形与这组潜在特质间的关系，可通过一条连续递增函数来加以诊释，这个函数便叫做项目特征曲线（Item Characteristic Curve，ICC）。任何一条项目特征曲线所代表的含义是：**答对某一试题的概率，是由考生的能力和试题的特性所共同决定。考生的潜在特质或能力的程度越强（或越高），其在某一试题上的正确反应概率便越大。**IRT 通过项目反应模型来确定考生的心理特质值和他们对于项目的答对或答错反应之间的关系，其数学表达式包括两类参数，即题目参数和能力参数。IRT 是建立在强假设基础上的，其基本假设为：①单维性假设，即考生的某一测验结果只取决于一种能力，其他能力的影响均可忽略；②局部独立性假设，即考生答该题时不受其他试题的影响；③单调性假设，即被试对题目所作出的反应概率遵循一定的函数关系。IRT 诞生后，在题库建设、学生水平监测、教学质量监控等领域中发挥越来越大的作用。

相较于 CTT，IRT 的优点主要是：①题目参数稳定，不受考生样本的影响，题目参数估计更为精确；②针对每个考生提供个别差异的测量误差指标，因而能精确推算考生的能力估计值；③解决了测验等值问题，它既能实现被试测验总分等值，又能实现题目参数等值；④定义了信息函数这一综合质量指标，来评定某个题目或整个测验的准确性。然而，IRT 也因为自身缺点而使其应用受到限制：①数学模型复杂，计算工作量大，单纯依靠手工很难完成；②估计参数时必须通过测试获得，被试不同则测验数据就不同。若要得出稳定的参数值，就要求测验项目和模型拟合，但拟合性指标依然严重依赖于被试样本的大小，样本过小则难以检测出数据与模型间存有的偏差。

三、概化理论（GT）

CTT 中的测验误差是一个笼统的概念，对于误差的来源和大小并未明确说明。为解决这个问题，20 世纪 70 年代初，克伦巴赫等人提出概化理论（Generalizability Theory，GT）。**GT 基于测验情境关系说，对 CTT 中过于笼统的误差进行探查和分解，通过辨明误差的不同来源，探究各个误差来源对测量结果的影响，进而优化测量过程，在可能的实施条件下优选实际测验方案，从而实现测量结果可靠性的最大化。**概化理论认为，测验情境关系由测量目标（object of measurement）和测量侧面（facet of measurement）两部分构成，其中，测量目标指测验真正想要测量的心理特质或能力，如阅读能力、数学能力等；测量侧面指影响测量目标的各种因素和条件，如测量工具、测量环境、评分者等。

基于 GT 的目的，具体分析过程由 G 研究和 D 研究两部分组成。G 研究是在观测全域上，根据测量设计对测量目标、测量侧面以及它们之间的交互作用的方差协方差分量进行估计，简而言之就是估计测验的各种误差来源和误差量。D 研究则是在 G 研究基础上，通过改变测量侧面结构、测验模型等，考察在不同测量情境下测量结果的稳定性和可靠性，确定最佳的测量方案，从而为有效控制误差、提高测验精度提供参考。

GT 的优点在于：①在理论假设上，与 CTT 的“严格平行测验假设”不同，GT 代之以“随

机平行测验假设”，使分析问题的条件较容易得到满足，适用范围具有广泛性；②利用方差分析技术，将测验误差分成几个部分，从而能辨明误差的来源；③主张要研究测量问题必须先确定测验情境，并在一定范围内变动测验的情境关系，以寻求最优化的测量设计，改进并指导实际的测量工作。GT 也存在一定局限：①注重所测潜在特质的单维性，但在实际的测量中却常常难以满足；②运用方差分析技术分解测量的误差来源，理论上成立，但计算繁琐，数据复杂。

四、认知诊断理论（CDT）

CDT 强调心理测量学与认知心理学的结合，**通过测验建立起观察分数和被试的内部认知特征之间的关系，进而对个体的认知过程、加工技能和知识结构开展诊断评估。**认知诊断充分吸收了认知心理学对人类认知加工过程的内在机制研究的丰富成果及独特研究范式，开发出了具有认知诊断功能的心理计量模型，即认知诊断模型。当前常用的认知诊断模型主要是基于项目反应理论框架并融入了认知变量的测量模型，如规则空间模型（RSM）、属性层级模型（AHM）、决定性输入噪声与门模型（DINA）、融合模型（FM）、多维项目反应理论模型（MIRT）等。

CDT 对学业成绩的评价超越了单独的分数评定或能力指标，能够挖掘深层信息，对学科能力做出具体解释，为新课程改革所强调的素质教育提供有力的评价和诊断工具，具有极大的应用价值。然而 CDT 至今未见广泛的实际应用，主要困难如下：①受制于认知心理学本身的发展，其研究结果很多是解释性、描述性的，在测验领域实际应用时可操作性不强；②开发出实用有效的认知诊断模型较为困难；③认知心理学和测量学两学科的沟通与结合不顺畅，复杂的数学模型令认知心理学家望而却步，而测量学家常常不能很好地驾御认知心理学必备的知识。

教育考试改革难以完全套用现有的教育测量学理论模型和照搬国外模式。为提升中国教育考试的科学化水平，要开展教育测量学理论和方法的深度研究，拓展已有测验模型，解决测验等值问题，制定教育考试数据评价标准，组建专家团队合力解决教育考试改革中的共性问题，充分发挥教育考试具有的鉴定、诊断、调控和改进提升的功能。

（供稿：刘芳、周怡、高柳萍）

【实践应用】

高中学业水平等级性考试数据分析拓展研究

沈励、万雅奇（北京教育考试院）

高中学业水平等级性考试（简称“等级考”）由于采用等级赋分的方式计入高考总分，并被用作高等学校录取的依据之一，因而兼具标准参照测验与常模参照测验的性质，这对高考试题命制和考试数据分析工作提出新要求。本文以某省某学科等级考适应性测试数据中的 2000 份随机样本为研究对象，拓展 CTT 和 IRT 两种理论应用于传统的等级考数据分析方法，通过 CTT 下的亚组分析及试题难度参数结构分析、多级计分 IRT 下的试题类别特征曲线和信息函数，分析试卷对不同能力考生的区分程度、试题难度结构配比、非选择题分值及采分点设置的适宜性，为考试命题工作及试题质量评价提供实践新思路。

一、CTT 框架下的分析模型

1.试卷对不同能力亚组考生的区分能力。按样本群体卷面总分从高到低排序，计算一分一段频数、百分比和累计百分比，再按某省赋分方案（21 个等级）中各等级比例要求，将样本数据归到不同亚组中，通过计算不同亚组卷面的难度系数，获得试卷对不同能力亚组考生的区分能力。结果表明，以中高端能力考生为例（前 10 个等级），亚组难度系数呈均匀递减趋势，与赋分方案每级间差 3 分的设计较为吻合，试卷对中高端能力考生的区分比较均匀。

2.试题难度参数结构表。将样本群体卷面总分相同的考生归为一个亚组，将该科目的 30 道试题的难度系数按 0.1 步长划分为 10 档，计

算每一个亚组中 10 档难度分类下的试题平均难度系数，获得试题难度参数结构表，用以分析整个试卷的试题难度结构配比问题。结果表明，通过检视 10 档试题难度在不同亚组上平均得分率的拐点位置、坡度变化，可以反馈当次等级考命题质量，对下一次试题命制时难度结构的预分析提供帮助。

二、IRT 框架下的分析模型

非选择题的命制需要设置合理的满分值和采分等级数（满分值加 1），等级过少则区分能力低，等级过多则增加评分误差。为研究非选择题满分值及采分点设置的适宜性，采用 IRT 中的多级计分 GRM 模型，以试题类别特征曲线和信息函数为基础，用以分析相邻两个得分等级的难度阈值对考生作出有效区分的程度，以及不同分值的非选择题提供的实际信息量与理论上应提供的信息量的匹配程度。

1.GRM 模型试题参数估计。采用 GRM 模型获得试题的难度和区分度指标，用以分析试题对考生能力的区分效果以及各采分点难度设置的合理性。当试题难度或区分度不够理想时，仍要结合专家判断、试题特征曲线、试题信息量等方式对试题质量做综合分析。

2.类别特征曲线图分析。等级性应答的试题特征曲线又被称为类别特征曲线，按照等级数量的不同，曲线就有多条。类别特征曲线图的横轴为考生的能力水平θ，纵轴为作答的正确概率 P(θ)，每一条曲线代表了不同能力水平的考生在该题中获得相应等级分数的概率。通过分析不同采分点等级的累积概率曲线面积是否能较好区分开来，以此表明采分点对不同能力考生水平的区分效果，用以判断该题满分值和采分点设置的合理性，如果不合理，可以考虑合并采分等级的方式对满分值进行调整。

3.试题信息函数图分析。试题信息量表示的是试题评价考生能力水平的准确性。试题信息函数图的横坐标为能力水平θ，纵坐标为试题信息量 I(θ)，每条曲线代表了该试题在不同的能力水平下所能提供的信息量值。信息量越大表示对考生水平的估计越准确。

4.试题信息量匹配分析。目的是通过构建所有试题在不同能力水平考生中的局部特征表达模式，以及分析不同分值的非选择题提供的实际信息量与理论上应提供的信息量的匹配性，用以检验命题预期和改进试题质量。具体分析方法有两步。第一步，计算试题最大信息量 $I(\theta)_{max}$ 及其对应的能力水平θ值，据此作散点图，得到每道试题对什么能力水平的考生具有最良好估计的直观判断。第二步，对试题理论上应提供的信息量比例和试题实际提供的最高信息量比例作对比。理论上应提供的信息量比例等于某道非选择题的满分值占所有非选择题满分值的百分比，实际提供的最高信息量比例等于某道非选择题信息量的最大值占所有非选择题信息量最大值的百分比。综合试题最大信息量分布及信息量匹配分析，可以透视试题命制方面的问题，提出改进试题的措施。

（摘编自《中国考试》2022 年第 5 期）

基于项目反应理论的研究生招生考试命题质量评价

宋学玲、梁正妍（教育部教育考试院、华南师范大学心理学院）

全国硕士研究生招生考试是国家高层次人才选拔的主渠道。对研究生招生考试初试的命题质量进行分析，探究其与测试目标群体的匹配程度是考试管理的重要环节。本文以 2022 年全国硕士研究生招生考试《心理学专业基础（312）》科目为例，以随机抽取的 22126 份作答数据为研究样本，运用 SPSS 和 R 软件，采用 IRT 中的双参数 Logistic 模型和拓广分部评分模型分别对选择题和主观题（简答题、综合题）试题进行参数估计，通过对试题质量参数及考生能力参数的分析来反映考试的整体质量，并通过信息函数对试题和试卷的测量精度进行探讨，以期拓宽研究生招生考试的命题质量评价路径，并为后续考试大纲的完善和命题质量的提高提供心理测量学上的参考。具体研究结果如下。

1.难度、区分度分析。《心理学专业基础》试卷共有 83 道题，其中选择题 75 道、简答题 5 道、综合题 3 道。就难度而言，难度参数值越高，试题难度越大。本套试卷中绝大多数试题难度都在[−3，3]之内，难度小于−0.5 的试题偏多。从主观题看，各题平均难度不超 0.5，大部分试题的各级难度也是负数多、正数少。可见此套试卷难度中等偏易。就区分度而言，区

分度参数值大于等于 1.5 为优级试题，[1，1.5)为良级试题，[0.5，1)为中级试题，小于 0.5 为差级试题。本套试卷优级和良级试题共有 43 道，占总题量的 51.81%。但是，试卷中仍有 10 道差级试题，中级试题占比也偏高，试题质量仍需改进。

2.考生能力水平分析。考生的能力范围分布较广，主要分布在[−2，2]范围内，其中能力水平在 0.5 左右的考生人数最多。

3.试题试卷信度分析。IRT 的信息函数对应 CTT 中的信度，反映考试分数对考生能力估计精度的指标，函数值越大，估计越精确。从项目信息函数来看，整套试卷中绝大多数试题的最大信息量都高于期望信息量，没有达到期望信息量仅 9 道题。从测验信息函数来看，对绝大多数考生而言，测验信息量都大于 10，对应 CTT 中的信度约为 0.9，测验信度高。但测验信息函数曲线整体偏左，信息量的最大值所对应的考生能力参数约为−0.8，可见，试卷对能力水平中等稍偏下的考生群体区分表现更好，对于优秀考生的筛选精度不够，客观上加大了复试的选拔压力。

4.知识板块分析。以往采用 CTT 或 Rasch 模型的相关研究均缺乏对知识板块层面的分析，本研究基于 IRT 做出了尝试。《心理学专业基础》试卷考查内容涵盖心理学导论、发展与教育心理学、实验心理学、心理统计与测量四个知识板块。研究表明，心理学导论、发展与教育心理学的试题在区分度表现上差于实验心理学、心理统计与测量。四个知识板块的平均最大信息量均远大于期望信息量，但心理学导论、发展与教育心理学实际提供的信息量比例低于应提供的信息量比例，而实验心理学、心理统计与测量实际提供的信息量比例则高于应提供的信息量比例。

（摘编自《心理与行为研究》2023 年第 2 期）

高中平面向量的认知诊断研究

秦海江、霍学晨、郭磊（贵阳市第三十七中学、厦门双十中学、西南大学）

传统教育测评以分数评价学生，认为分数相同的学生能力水平也相同。然而，分数相同并不代表认知结构和知识掌握情况相同，简单以分数高低为依据的教育评价不尽合理。《深化新时代教育评价改革总体方案》提出坚决克服唯分数、唯升学等顽疾。认知诊断评估作为最新一代的智慧测评技术，能够直接探索学生的潜在知识状态并给予认知结构层面的评价与反馈，是当下解决传统教育评价中唯分数论等弊端的重要突破之一。本文以高中平面向量知识为例，基于认知诊断评估技术、认知诊断测验主流的编制方法，编制关于“高中数学平面向量”的诊断性测验，并以 X 中学的 1363 名高一年级学生为研究对象，展示认知诊断评估技术的实例分析，以帮助教育者从学生的认知结构层面讨论对相应知识点的掌握情况，指导后续开展的针对性补救教学。

一、测验编制

通过参考相关资料和教师交流讨论，将平面向量知识划分为 4 个属性：线性运算 A1、数量积 A2、坐标 A3、综合运用能力 A4，进一步研究确定属性间层级关系为独立型，并以此构建形成 Q 矩阵。在此基础上，通过采用改编的高考题目和各省模拟试题，形成 10 道测验题目，邀请 5 位高中数学老师对所有题目的 q 向量进行编码，计算它们标定的一致性（肯德尔 w 系数），得 w=0.937，P<0.001，表明构建的 Q 矩阵一致性程度高。

为验证编制测验的质量是否合格，对 1363 名学生进行施测，获得试题和测验的难度、区分度、信度、HCI 等指标。研究结果表明，各小题的难度与区分度均符合测验编制的基本要求。并且整体测验的难度为 0.56（难度适中），区分度为 0.49（区分度良好）。用层级一致性指标（HCI）验证 Q 矩阵属性层级结构的合理性。一般认为 HCI 在 0.7 以上即可认为拟合良好，本研究所得 HCI 为 0.777。

二、应用实例

采用认知诊断模型中的 G-DINA 模型对 1363 名学生的认知结构进行探索与评估。计算由 G-DINA 模型估计得到的学生属性总体掌握情况与学生原始卷面总分的相关系数，相关系数为 0.859，相关程度高，表明认知诊断评估的结果对原始分数的解释程度高。认知诊断评估

既可从群体角度描述群体评价，也可以对个体进行详细的评价，具体如下。

1.群体诊断反馈。92.2%的学生被划分到8种知识状态中。对X中学总体、不同校区（A和B）和不同等级班级的学生的属性掌握情况进行分析。结果表明，总体上，X的高一学生在刚完成平面向量的教学时对各属性的掌握情况并不理想，掌握最好的属性A2的概率仅61.2%，属性A1最低（40.0%）。在两个校区中，B校区的学生掌握各属性的概率均要高于A校区。在不同等级班级中，总体看来，属性掌握概率随着班级等级的降低而依次下降，与行政班级的等价划分相一致。进一步通过ANOVA方法验证上述认知诊断结果的有效性，以A、B两校区在掌握A1属性上的方差分析结果为例：$F(1)=130.63$，$P<0.001$，$\eta_p^2=0.088$，表明B校区的学生在A1属性的掌握概率上显著高于A校区，与上述结论一致。

2.个体诊断反馈。通过向每一位学生反馈卷面分数、知识状态、以及属性掌握概率，可以发现每一位学生对各属性的掌握情况，具体有哪些不足并针对性地改进。也可以发现不同学生虽然取得相同的卷面分，但认知结构上仍然存在区别。还可以发现学生卷面分高不代表能力水平一定高，也有一些属性需要进一步巩固。

3.学习进阶路线。将认知诊断分析与项目反应理论相结合，针对不同知识状态的学生制定精细而具体的学习进阶路径。该路径以能力参数为基础，建立各个知识状态之间的联系：某一种知识状态下对应的能力参数值越高，则说明该状态越难以达到。学生可以通过学习未掌握的属性知识，不断向其他较为容易达到的知识状态转变，最终掌握所有属性。教育者可以根据达到每个知识状态的难易程度决定教学顺序。

（摘编自《数学教育学报》2024年第2期）

基于多元概化理论的中小学教师资格考试质量分析

——以《综合素质》（中学）科目为例

杨宏博、赵轩（教育部考试中心）

教师资格考试是衡量教师资格制度现代化和科学化的重要指标，是国家进行教师资格制度改革实践的突破口。随着中小学教师资格考试的快速发展，检验教师资格考试改革的有效性，检测考试评价本身的科学性和公平性，建立科学有效的考试质量评价方式成为重要的研究课题。本研究以中小学教师资格考试笔试科目《综合素质》（中学）试卷为例，以随机抽取的2372份作答数据为研究对象，采用mGENOVA软件，运用多元概化理论探讨试卷各模块及全卷的测量信度，考查内容模块样本容量变化对考试信度的影响，分析各模块对总测验的贡献率，为试卷优化、提高命题质量提出合理化建议。具体研究结果如下。

一、G研究结果

本研究将试卷分为职业理念、教育法律法规、教师职业道德规范、文化素养、基本能力、写作能力等六大内容模块。G研究采用$p\times i$六内容因子随机单面交叉设计，得到各效应在六个因子上的方差和协方差变量估计矩阵。结果表明，教育法律法规和文化素养模块与其他因子的协方差分量较小，说明这两个模块得分高低顺序与他们在其他模块中的顺序不太一致，即这两个模块中的题目在区分考生能力方面功能较弱。在效应p上，写作能力因子的方差分量最大，表明对考生的区分能力较强，教育法律法规和教师职业道德规范因子的方差分量较低，表明对考生的区分能力相对较弱。

二、D研究结果

一是各因子全域分数估计精度。D研究采用$p\times i$六内容因子随机单面交叉设计，基于G研究估计的方差与协方差矩阵，进一步估计出考生在六个因子上的全域分数及相应误差项的方差分量，进而估计概化系数与可靠性指数。结果表明，六因子中全域分数方差分量从高到低依次为写作能力、基本能力、文化素养、职业理念、教育法律法规、教师职业道德规范。考虑到各分量误差方差的因素，测量信度最高

的因子是写作能力（可靠性指数为 0.806），表明写作能力模块的测量信度较高。

二是全域合成分数的测量精度。本研究按照各测量分项试题量所占比例来确定权系数 b，职业理念、教育法律法规、教师职业道德规范、文化素养、基本能力、写作能力这六个因子的权系数分别是：0.147、0.235、0.147、0.265、0.147、0.059。对六因子全域分数进行合成，得到全域总分的方差为 0.090，全域合成分数相对误差方差为 0.037，全域合成分数绝对误差方差为 0.106，进而计算出全域合成分数的概化系数为 0.707，可靠性系数为 0.458。可见，全域合成分数的概化系数较高。而六因子未进行全域分数合成时，各因子全域分数的概化系数及可靠性系数均较低，在全域分数合成总分后测量精度显著提高（除写作能力因子外），因此对六个分测验的分数进行合成是合理的。

三是各因子对总方差的贡献比例。通过估计各模块对考试总分方差的实际贡献率（比例），得到六个模块对试卷总分方差的实际影响程度。结果表明，写作能力因子对总体方差的贡献比例较试卷赋分比例高，其它因子对总体方差贡献的比例较试卷赋分比例略低。总体而言，各分测验基本达到考试的预期测量目的。

四是各因子样本容量对测量信度的影响。为改善测验方法，进一步改进测量信度，本研究考察了各因子样本容量变化对各分测验自身及试卷总分测量信度（采用总分概化系数作为信度指标）的影响情况。结果表明，当各因子题量为 2 倍模式时，全域总分的概化系数可增至 0.828；当各因子题量为 3 倍模式时，全域总分的概化系数可增至 0.878。由于职业理念、教育法律法规、教师职业道德规范、文化素养、基本能力因子的全域分数误差方差相对较小，因此，提升这几部分的题量对整卷的概化系数的影响并不显著。而写作能力因子容量提升至 2 倍时，全域总分的概化系数可增至 0.841，提升至 3 倍时，全域总分的概化系数可增至 0.897。可见其对测量信度影响显著。不过，综合考虑考试性质、试卷长度、作答时间限制等因素，通过增加写作题的题量以提高测量信度的方案在实际操作中不太可能实现，在 120 分钟考试时间内，保持现有的试题数量是比较合适的。

（摘编自《心理与行为研究》2019 年第 2 期）

全国英语等级考试锚测验非等组设计中样本量对等值结果的影响

景春丽、马洁、章建石（教育部考试中心）

全国英语等级考试（简称 PETS）是教育部考试中心设计并负责的全国性英语水平考试体系。除 PETS-4 外，其他级别的考试每半年举行一次，采用锚测验非等组设计进行等值，在每次正式考试前一周左右，随机抽取 300 名左右参加本次考试的考生参加锚测试。很多时候由于抽样的限制，样本量往往达不到 300 人，而样本量是影响随机误差最直接的指标之一，在这种情况下，等值结果是否精确？基于此，本研究采用某次全国英语等级考试（PETS-5）的锚测验数据和实测数据，从参加锚测验的 660 名考生样本中随机抽取 30、60、90、120、150、180、210、240、300 人作为样本，再从正式考试中抽取 10000 人（包括随机抽取参加锚测验的样本），通过考号将锚测验和正式考试的成绩链接起来，基于 Rasch 模型进行参数估计，采用以 Bigsteps 为核心的自主改进软件进行参数估计和参数转换，探究不同锚测验样本量对 PETS 等值结果及其稳定性的影响。具体研究结果如下。

1.抽样的合理性分析。基于经典测量理论验证抽样是否合理。对不同样本量锚测验的观察分数进行方差分析和单样本 *t* 检验，结果表明，不同样本量锚测验观察分数的均值不存在显著差异，且不同样本量的抽样与实际考生样本之间锚测验的观察分数也均不存在显著差异，综上可知，研究的随机抽样合理。

2.参数估计。在锚测验非等组设计中，不同版本测验中项目参数和能力参数的转换均是通过锚题参数的平均值和标准差实现的，而 Rasch 模型在锚题参数转换过程中只用到了锚题参数均值。从不同样本量锚题难度参数均值可以看出，不同样本量锚题难度参数均值为 -0.09~-0.04，随着样本量的变大，难度均值变

化越小，在样本量达到 150 之后，参数均值趋于稳定。从不同样本量锚题难度与给定锚题难度之间的相关系数也可以看出，随着样本量的变大，相关系数趋于稳定，在样本量达到 150 以后，相关系数稳定在 0.7~0.72。

3.估计差异分析。本研究从两个方面考虑参数及等值差异：一是考虑样本量不同时锚测验所估计出的锚题难度值与给定锚题难度值之间的差异；二是考虑不同样本量锚测验对应的试卷等值结果的差异。同时，以样本量为 660 的锚测验等值结果为标准，比较不同样本量的锚测验的等值结果与样本量为 660 的锚测验的等值结果。计算差异的指标为均方根离差（RMSD）。结果表明，不同样本量估计出的锚题难度值与给定锚题难度值的差异较小，但是当样本量介于 30 到 120 之间时，RMSD 值并不稳定，当样本量达到 150 之后，RMSD 值趋于稳定。并且不同样本量锚测验对应的试卷等值结果与设置的标准之间的均方根离差较小，但是当样本量介于 30 到 120 之间时，RMSD 值并不稳定，当样本量达到 150 之后，RMSD 值趋于稳定。

4.等值结果对实际考试结果的影响。垂直量表是将测量领域相似但考查的内容水平不同的数个测试构建到一个共同量表上的过程，即在测试内容相同但水平不同的测试之间，通过共同量表，使得试题的难度或考生的水平能够在数值上相互比较。本研究用不同样本量锚测验将实际考试题目参数转换到给定锚题的量表上，通过自主研制的计算能力值软件（abli）计算合格能力值对应的客观题实际分数线。结果表明，用不同样本量锚测验得出的实际分数线是 47 或者 48，但是当样本量达到 150 以上，实际分数线就稳定在 47。

总之，对于目前的 PETS-5，考前进行的锚测验样本量确定在 150 以上就可以得到比较稳定的结果。但需要注意的是，试题参数估计与等值试卷的长度、题型及试题的性质有关。当这些因素变化时，对锚测验样本量的要求也可能会发生变化。因此，一个考试要采用多大的锚测验样本量，要具体分析，不能一概而论。

（摘编自《中国考试》2017 年第 6 期）

【科研动态】

全省第十七期教育考试机构专业化能力提升研修班顺利举办

8 月 19 日—23 日，全省第十七期教育考试机构专业化能力提升（教育测量与评价专题）研修班在江西南昌顺利举办。来自全省教育考试系统及部分高校的 35 名学员参加培训。省教育考试院副院长吴成兵参加开班仪式并作开班动员讲话。

吴成兵强调，教育考试评价是教育科技人才一体化战略的重大组成部分，科学反映学科核心素养、精准衡量学业质量标准、深入推进教考衔接、助力全面育人，是教学领域、考试领域、评价领域共同的现实任务。要克服教育考试评价可解释化的短板，实现科学化的考试评价，建立以信息技术支撑的卷库题库、网络考试，离不开现代教育测量理论的指导和运用。学习研究和深入运用现代教育测量理论和技术，是教育考试工作者的必然选择和自我追求。

本次研修班特邀了江西师范大学心理学院在国内教育测量与评价领域知名的专家团队授课，课程内容聚焦现代教育测量理论的核心，围绕项目反应理论、计算机自适应测验、认知诊断测验、教育增值评价等内容展开，不仅提供了高屋建瓴的理论分析，还涵盖了详实可行的实践操作指南。参与此次培训的学员们纷纷表示收获颇丰。培训课程的安排紧凑而高效，内容既深入又实用，紧密贴合当前教育评价改革的实际需求，具有较强的实践价值和指导作用。

【成果展示】

高中学业水平考试试题难度模型建构研究

刘芳　王伟群　吴星

（江苏省教育考试院、苏州大学材料与化学化工学部、扬州大学化学化工学院）

在国家新一轮高考改革中，高中学业水平考试成绩已明确作为学生毕业和升学的重要依据。经典测量理论（CTT）下的试题试卷难度易于理解。本研究在明确影响试题难度主要因素的基础上，通过编写测试题进行抽样施测，依据难度统计数据建立影响因素不同呈现类型的难度序列及赋值规则，最终建构了一种具有较高信度和效度的试题难度分析模型。

一、确定影响试题难度的主要因素

在中国知网上通过主题词查询对有关试题难度研究的文献进行了检索和筛选，并对文献进行质化分析，得到影响理科试题难度的主要因素有 11 个，分别为学科知识综合度、情境陌生程度、知识综合运用程度、信息复杂程度、知识点难易程度、解题步骤多少、结果呈现方式、信息量多少、信息呈现方式、答案猜测程度、知识记忆难易程度。

采取问卷调查的方式收集数据，探究 11 个因素之间存在的相互关系。选用 2016 年、2017 年高考江苏化学卷的非选择题编制问卷。每份问卷包含试题部分和选填影响解题制约因素部分。对调查结果的统计数据样本开展主成分分析，11 个难度影响因素被降维归属为 3 个主成分，与试题情境、问题解决过程、试题答案相关，这与问题解决理论的三段论相一致。

在问题解决的逻辑体系下，通过对命题专家、一线教师的访谈，结合对高中生的解题思路访谈和答题出错分析，从问题表征、问题解决和结果输出 3 个维度上，重新梳理影响试题难度的主要因素，从而确定信息呈现方式、信息利用方式、情境陌生程度、知识综合程度、思维层次、答案表达形式、答案开放程度等 7 个因素为影响高中学业水平考试试题难度的主要因素。

二、建立试题难度影响因素的赋值规则

建模之前必须对每一个难度影响因素赋值。由于每一个影响因素在实际试题存在多种呈现类型，呈现类型不同，考生作答试题所感受的难度是不一样的。以化学学科为例，列举每一个因素在高中学业水平考试试题中常见的 3 种类型。当评判 7 个主要因素的不同呈现类型对试题难度的影响时，发现知识综合程度、思维层次、情境陌生程度、信息利用方式等因素的不同呈现类型对试题难易变化趋势的影响是明确的，即知识越综合，试题难度越大；思维层次越高，试题难度越大；情境越陌生，试题难度越大；信息越隐蔽、干扰性越强，试题难度越大。而信息呈现方式、答案表达形式和答案开放程度等因素的不同呈现类型，对试题难易变化趋势的影响则不是很清晰，如信息呈现方式的纯文字、含图表、含流程 3 种类型。因此，通过编制测试题，并在全省范围进行抽样测试，根据实测的 CTT 难度对这几个因素不同呈现类型的难度序列进行评判。

利用专家评定法和实证研究方法厘清了 7 个因素的不同呈现类型对试题难易程度的影响。在此基础上，形成同一因素不同呈现类型的难度序列，并据此进行分级赋值。考虑到最终通过难度分析模型得到的结果为难度系数，难度系数越小，则试题难度越大。因此，按照由难到易的顺序，分别以 1、2、3 对同一因素不同呈现类型进行赋值。

三、建构试题难度分析模型

选取 6 位具有丰富教学和命题经验的中学化学教师，进行赋值规则培训后，让他们独立对所有测试题的每一个设问从 7 个影响因素的维度进行赋值。对教师的赋值结果开展一致性分析，得到肯德尔和谐系数为 0.845，表明 6 位教师的赋值在 0.01 水平上是一致的。在此基础上，对个别有差异的赋值分析原因并调整赋值，最终形成所有测试题的结构赋值。

在基于机器学习的建模过程中，随机选取全部数据的 70%作为训练集，建立线性回归方程，再以剩余的 30%数据作为测试集进行模型验证，得到难度系数误差的平均数和误差的标准差，以检验所建模型的准确性。为了在有限数据的情况下尽量作出拟合度较好的模型，实际操作时进行了多次建模，并对多次建模结果从减少偶然误差、参数实际解释意义等多个角度筛选，最终获得基于经典测量理论（CTT）的难度分析模型：$y = -0.4281 + 0.0820x_1 + 0.1082x_2 + 0.0028x_3 + 0.0952x_4 + 0.0509x_5 + 0.1360x_6 + 0.0102x_7$，这里 y 为试题 CTT 难度系数，x 为难度影响因素的赋值。利用模型计算所有测试题的预测难度系数，并与相应测试题

的实测难度系数进行对比，结果表明，测试题难度的模型预测值与实测值有较好的拟合度。

为进一步验证所建模型的实用性，选取2015年、2016年和2017年高考江苏化学卷中12道非选择题，在对每一道试题从7个影响因素维度赋值后，利用所得模型预测数据与相应试题的实测数据进行对比。检验结果表明，模型预测数据与实测数据有较好的拟合度，建构的模型具有一定的应用价值和推广意义。

（摘编自《化学教育》2022年第21期）

改善新高考学考评价的教育测量学探索

黄红波（江苏省教育考试院）

本研究以2018、2019两年某科目学业水平考试的选择题数据为研究对象，采用简单不重复随机抽样法，分别抽取各1000名考生。基于经典测量理论（CTT）和项目反应理论（IRT）统计难度、信度/信息量、区分度等指标，并从试卷、试题和考生三个层面分析了IRT相较CTT应用于新高考学考的优势，探索将考生水平和试卷难度参照到同一能力量尺上的测量学方法。

一、研究结果与结论

从试卷层面来看，CTT分析表明，两年的考试难度都比较容易，信度高，区分度一般，尤其2019年。IRT分析表明，2018年试题分布较为分散、信息量分布与人数分布较匹配，2019年试题集中在低水平区域、信息量分布狭窄，缺乏高水平能力试题。从试题层面来看，2019年学考的CTT和IRT试题难度总体上非常一致，但个别试题存在差异显著，主要是CTT下因素“人”和“题”的相互依赖所致。从考生层面来看，考生的CTT分数和IRT能力在总体趋势上基本一致，但它们之间不满足简单线性关系，显示出CTT分数不能满足分数的等距性。

研究表明，用项目反应理论（IRT）的三参数模型（3PLM）分析学考项目具有很好的拟合优度，可以进一步探索推广到其它学考项目。IRT在潜在特质理论基础上建立数学模型，量尺具有等距特征，突破了传统CTT分数不能刻度能力的局限。而且，IRT拟合后的数据结果，具有样本独立性和测验独立性，可以解决CTT分数中考生能力和试题难度相互依赖难以区分的困难。将IRT应用于学业水平考试，可以独立地从考生水平、试题难度，以及两者匹配关系等方面，更加科学深入地评价考试命题和学业质量。

二、对考试评价的启示

考试分数一直都是最重要的教育评价证据，选择好与评价目的一致的分数量尺对做出科学评价至关重要。IRT技术通过项目反应曲线数学模型的拟合，将每个考生对每道试题的反应量化为分数，把“人”和“题”刻度在同一能力量尺上，实现能力值等距，是科学准确地反映能力内涵的有效工具。

1.试卷评价从要求“合适”到追求“合理”。难度是试卷评价的重要指标，一份高质量的试卷必然要求难度合适。但CTT试卷评价中的试题试卷难度不能有效反映各类难度试题与考生能力水平的对应关系，因而CTT分析下的难度合适未必是一份好卷，如上述提到的2019年学考卷。而在IRT分析下的情况有所不同，考生能力和试题难度表征在同一能力量尺上，能够呈现各能力水平所对应的试题分布，可以更精准地评价试卷质量。只有当试题难度与考生水平“合理”匹配，才能保证考试的覆盖度，达到试卷较高的效度和区分度。

2.试题评价从依靠“经验”到依据“实证”。目前，试题评价大多数采用专家法，试题质量的优劣主要依靠学科专家的经验判断。学科专家尽管学科经验丰富，但常缺乏教育测量研究，容易把价值判断重点放在学科知识上，对试题实际启动考生的认知操作认识不够深入，导致经验判断和实测结果经常会产生偏差。而IRT具有样本独立性，试题参数直接反映试题内在特征，是试题质量评价的可靠“实证”。依靠IRT技术对试题难度、区分度和信息量等参数进行标定，专家据此对试题能够做出更加客观、准确的评价。

3.学业质量评价从显示“排位”到揭示“内涵”。IRT的测验独立性为准确认识考生水平提供了可能，从而比较确定地回答究竟是试卷难还是考生差、试卷容易还是考生进步的问题。将IRT技术应用于学考数据分析，一方面，有利于引导学校正确认识考试成绩，基于学生能力形成来诊断并改进教学，防止教育评价的过度攀比和片面追求高分的乱象。另一方面，有利于考试结果从能力内涵更好地反馈命题质量，确保学考命题对标学业质量标准，将高考、学考和教育教学质量监测在能力衡量上保持一致。

（摘编自《教育与考试》2021年第3期）

【科研讲堂】

文献综述“找不到文献”的三个解决策略

胡乐浩

文献综述是学术研究中不可或缺的一环，它对于理解特定研究领域的背景、现状和未来趋势至关重要。通过系统地回顾和分析现有的文献，作者可以发现当前研究存在的不足之处，从而提出自己的研究问题。而在这一过程中，文献梳理必不可少，它需要从现有的研究文献中找到与自己研究主题相关的内容，并对其筛选、分类和分析。通过这一过程，作者能够深入了解各个研究之间的关系，识别研究中的空白点和争议点，从而为后面的研究批评与研究问题的提出打下基础。不过，在文献检索这一环节，很多作者由于选题过于新颖或是检索方法生疏，而出现找不到或是只能找到少量与自己选题相关的文献，导致后面的文献批评只能基于倒推的逻辑展开。但倒推的使用虽然简单，要对其进行合理解释则很难，如此一来，很容易使论文在逻辑上暴露出缺陷。因此，本文将基于三种文献检索方法，以帮助大家解决在文献梳理环节找不到文献的难题。

一、从无到有推导型：A=B+C

很多作者在确定自己的选题后，首先习惯性要做的便是，将整个选题或题目名称直接置入文献数据库中，点击“检索”一番，结果喜忧参半。喜的是，“几乎没有人做过类似的研究，我的论文创新性很大”；忧的是，“没有与我相类似的研究，我的文献综述该怎么展开？”。其实，作者忧愁的这一方面，反映了文献综述逻辑的理解误区。文献综述环节是要还原研究问题的推导提出过程，它是一个“从无到有”的过程。若直接检索选题，那是一套反着的逻辑，必然会出现找不到文献的情况，因为研究问题本身即是创新的。即便是能够找到文献，也会因预设了“问题”，导致文献批评的过程过于刻意、机械。

在基于上述内容明确了文献综述的存在逻辑后，再来看文献检索环节。从现实角度来说，创新有两种来源：一是来自全新的想法或突破；二是在现有知识或技术的基础上进行组合、改进或融合，从而实现创新。但对于大多数人而言，尤其是在人文社会科学领域，要想实现完完全全的创新，几乎不可能。因此就要从第二个角度去思考，即研究问题是在现有知识组合的基础上实现创新的。按此逻辑，**如果我们将自己的研究问题看做 A，那么在文献梳理环节，我们就要去寻找能够基于“B+C”组合，去推导出 A 的内容**。下面以《ESG 评级不确定性对企业绿色创新的影响研究》（《管理学报，2024 年网络首发》）这篇论文来详细介绍这一文献梳理的操作逻辑。

这篇文章的研究问题为“ESG 评级不确定性对企业绿色创新的影响”，倘若我们直接拿着这一题目名称进行检索，会发现除了这篇文章之外，再无其他检索结果。即便是将这一题目中的关键词进行拆分组合检索，如“ESG 评级”“ESG 评级不确定性”“企业绿色创新”，也会发现高质量期刊论文文献寥寥无几。而该文作者则是将“企业绿色创新”“ESG 评级”分别梳理，因为关于这两者的文献非常多，最终推导出了该文的研究问题。其中，在“企业绿色创新”这一维度的文献梳理上，作者梳理了“企业绿色创新的影响因素”，而在这些内容里，“ESG 评级”位列其中。接着在“ESG 评级”这一维度的文献梳理中，作者的文献梳理逻辑为：ESG 评级结果—ESG 评级不确定性—ESG 评级不确定性对企业的影响，并最终推导出“ESG 评级不确定性很可能成为企业可持续发展的严重桎梏”。最后，由“企业可持续发展”落到了“企业绿色创新”视角上，并给出两个明确的理由，而这些理由都是与前文呈现的文献紧密相关、前后呼应的。进而，推导出了“ESG 评级不确定性与企业绿色创新的关系究竟如何？”这一研究问题。

从上一过程我们可以看到，如果把 B 当作一个圆，C 也是一个圆，那么 A 就是 B 和 C 两个圆交汇的部分，也就是说 A 是存在于 B 和 C 中的，是取它们的交集得出的。因此，我们在

围绕A检索文献时，一方面需同时检索B与C，另一方面需从中筛选“+”类文献，即两者存在交叉的研究，进而推导出A，而关于A的直接文献是不能够直接检索到的。

二、概念扩展延伸型：泛化检索

当选题中的主题词专指度过小时，也会遇到检索不到文献的情况。这个时候，我们就**需要对主题词做概念界定或是向上延伸，寻找它的上位词，做泛化处理。通过对主题词的泛化处理，能够减少搜索词的特定性，来增加检索结果的数量，从而筛选和评估出最相关的文献。**例如，苹果的上位词是水果，新质生产力的上位词是生产力。其原理为，泛化的主题词通常比具体的主题词覆盖更广泛的主题和概念，而且不同作者可能会使用不同的术语来描述相似的概念。此外，泛化的主题词可能包括了与原检索主题词相关的同义词、近义词或相关术语，这也能够有助于搜索系统找到使用不同表述但内容相关的文献。基于这一检索逻辑的不同，有如下两种检索策略：

第一种对选题中专指度过小的主题词做界定，找出已有研究中与之相类似的较为泛化的主题词，而后直接围绕后者做文献梳理。例如《破碎的自我：“小镇做题家”的身份建构困境》（《中国青年研究》，2021年第7期），该文发表于2021年，但知网检索显示，以“小镇做题家”为主题的学术研究在2020年仅有1篇，这是因为该词在彼时刚被提出不久。在这种已有研究极度缺乏的情况下，该文作者选择了与“小镇做题家”相似的群体进行了文献梳理，并阐释了这样做的合理性——“‘小镇做题家’作为一个网络新词，现有的直接相关的学术研究较少，所以本文从‘小镇做题家’群体的相似符号特征来综述研究现状。一方面是关于出身农村、进入城市的大学生群体的研究，另一方面是高学历获得者与身份建构的研究”。

第二种是“由大到小”，即直接从宽泛的角度做文献梳理，经一步步缩小，落到既定选题中专指度较小的主题词上。以《控制权理论视角下乡镇干部职业倦怠现象及其治理——基于江西省YF县FX镇乡镇干部绩效考核工作试点的调研》（《求实》，2020年第5期）为例，该文选题中主题词为“乡镇干部职业倦怠”，但在彼时之前，关于“乡镇干部职业倦怠”或是“基层干部职业倦怠”的研究寥寥无几，即便是以句子检索，高质量学术研究（核心期刊）文献也不足5篇。在此情况下，该文作者按照“职业倦怠—公务员职业倦怠—基层干部职业倦怠—乡镇干部职业倦怠”的逻辑梳理出了论文所需的已有研究文献。之所以如此可行，是因为“干部职业倦怠”是大环境，“乡镇干部职业倦怠”是大环境中出现的个例，所以我们可以基于“干部职业倦怠”的文献梳理来一步步去为文献批评与研究问题的推导做铺垫。

三、技术导向依赖型：句子检索

除了基于以上两种逻辑推理的思路去做文献梳理，知网等数据库中的“句子检索”功能也能在很大程度上解决检索不到文献的情况。它是一种更为精细的检索方式，允许作者通过输入两个或以上的检索词，在全文范围内查找同时包含这些词的句子。**这种检索方式特别适用于寻找包含特定概念或关系描述的句子，从而帮助用户更准确地定位到所需要的相关文献。**因为句子检索不仅限于篇名、关键词或摘要，而是可以在文献的全文范围内进行，这增加了检索的灵活性和可能性。而有时候，即使文献的标题或摘要没有直接包含检索的主题词，文献的全文中也可能包含深入讨论相关主题的句子，而句子检索便可以帮助作者发现这些内容。

仍以上面提到的“基层干部职业倦怠”这一主题词为例，在“来源类别”限定相同的情况下，常规高级检索只能检索出12篇文献，而在“句子检索”的模式下，能够检索出40条文献。通过这些直接检索出的文献，作者也可以发现不同文献间在论述上的联系，即使这些文献并未直接引用彼此，但对于作者的文献梳理或理论讨论部分的写作尤为有用，甚至可以直接将检索来的内容用作写作的引用内容，提高写作效率。不过，需要注意的是，句子检索对文献的覆盖面和数量并没有直接影响。如果数据库中相关文献本身就较少，或者检索词过于特殊导致没有匹配的句子，句子检索可能也无法返回更多结果。在这种情况下，作者便需要调整检索策略，比如更换或增加检索词，或者使用其他类型的检索方式来辅助寻找文献。

（本文选自“科研写作研究所”公众号）