基于教育测量理论的中学数学试卷质量评价研究

何　穗　吴慧萍

【摘　要】本文基于经典测量理论（CTT）和项目反应理论（IRT），对湖北省某地区的高中数学试卷是否符合命题规则和考核目标进行质性分析，同时根据考生的作答情况对试题进行量化分析。结果表明，使用IRT中的能力参数能更加全面真实地反映学生的能力水平，教师在教学过程中对学生应用能力的培养不够。

【关键词】试卷质量分析；教育测量与评价；经典测量理论；项目反应理论
【中图分类号】G40-058.1　【文献标识码】A　【文章编号】1674-1536（2012）08-0049-05

一、引言

《国家中长期教育改革和发展规划纲要（2010–2020年）》，明确把提高教育质量作为教育改革发展的核心任务，并多次强调与教育质量的监测和评价相关的内容。[1]很多国家对教育质量评价表现出浓厚的兴趣，影响较为广泛的教育监测项目有：国际数学与科学研究趋势（TIMSS）、国际学生评价项目（PISA）和美国国家教育进展评估（NAEP）等。由此可见，在全面提高教育质量、实施素质教育的今天，建立健全的教育质量监测体系已成为教育改革和发展的迫切需求。

在中学数学教学的实践中，教师分析试题或试卷的质量时，大多凭借自己的经验笼统地说好或不好，很少去进行统计、分析、测量、评价。本文利用经典测量理论（Classical Test Theory，简称CTT）与项目反应理论（Item Response Theory，简称IRT），尝试在中学数学试卷质量分析上作些深入研究。通过实证研究和量化结果了解学生对知识、技能的掌握情况，反馈教学中存在的问题，不仅能为优化和改善后续教学工作指明努力的方向，而且可以为日后修改或筛选考试试题、建立试题库和实施标准化考试等提供参考。

二、方法与过程

1.研究思路

本研究基于CTT和IRT两大教育测量学理论，结合统计学的有关理论和方法，通过对某数学考试卷进行研究，科学地分析试卷质量和考试结果，以研究教育测量技术在中学数学试卷质量评价中的实现方法，为判断教育活动现实的（已经取得的）或潜在的（还未取得，但有可能取得的）价值[2]提供数学教育测量的工具和参考模式。其技术路线如图1所示：

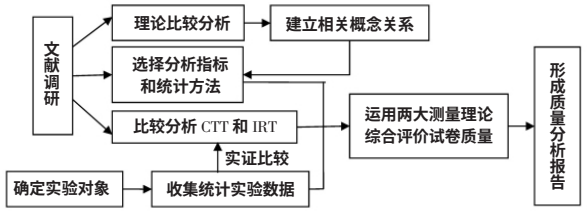

图1　研究技术路线

2.分析对象

为保证调研对象的代表性和模型的拓广性，本研究选取湖北省某地区某次比较大型的联考测试中的高一数学卷。同时，根据经典测量理论中抽样要有代表性和项目反应理论中样本量要大、被试范围要广等要求，随机抽取了3所参加联考的学校的考生作为样本，共1628名考生参与实验。

为尊重测试对象的隐私，本研究将测试卷命名为A卷。A卷采用网上批改，阅卷系统自动记录每位考生在各题上的得分，大大方便了样本数据的收集与抽取。A卷的测试范围为高中数学必修1的第一章到第三章，有10道选择题（1~10题）、5道填空题（11~15题）和6道解答题（16~21题），考试时间为两小时。

3.分析依据

（1）经典测量理论

试卷作为教育测量的重要工具，要达到良好的测试效果，必须做到真实可靠、准确有效、难易适中、

本文获湖北省教育厅科研项目《义务教育均衡发展监测评价研究》（W20120501）资助。
何　穗/华中师范大学数学与统计学学院教授，博士。（武汉　430079）
吴慧萍/澳门大学教育学院博士生。

鉴别力强，这也正是经典测量理论中的信度、效度、难度和区分度四个测验质量指标。[3]

经典测量理论又称为真分数理论，假定观察分数 X 与真分数 T 线性相关，即 CTT 的数学模型为

$$X=T+E$$

这里，随机误差 E 服从均值为零的正态分布。

经典测量理论虽然简单易懂，对心理与教育测量和实践的贡献巨大，但由于理论体系的先天不足，存在着不少的局限性，[4]如严重依赖样本、信度估计精度不高、难度和被试水平没有定义在同一参照系上，等等。另外，CTT 也无法回答总分相同的考生的真实能力有无差异等问题。

⑵ 项目反应理论

项目反应理论建立在分析和克服经典测量理论的缺陷的基础上，是一种新兴的心理与教育测量理论。与 CTT 的弱假设相比，IRT 的前提假设非常严格，主要包括单维性假设和局部独立性假设。[5]实践中应用得比较广泛的是伯恩鲍姆（Birnbaum，1968）的逻辑斯蒂克（Logistic）三参数模型（3PLM），其形式如下

$$P(\theta)=c+\frac{1-c}{1+e^{-Da(\theta-b)}},\theta\in(-\infty,+\infty)$$

参数 b 为难度参数，a 为区分度参数，c 为猜测度参数。其项目特征曲线如图 2 所示：

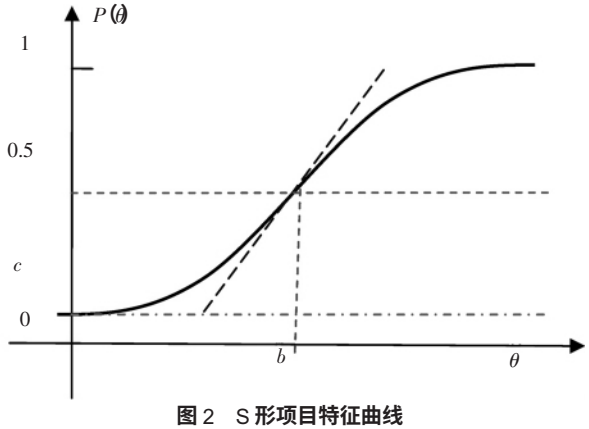

图 2　S 形项目特征曲线

项目反应理论在题目参数（难度、区分度等）的基础上，还提出试题信息函数[6]的概念，从而能根据反映测试要求与项目性能的信息函数曲线来挑选试题，提高题库参数的完备性和题库管理的可控性，为测验编制提供清晰的科学逻辑和强有力的指导作用。

应用项目反应理论指导测验编制和试卷评价时，必须对项目参数和特质水平（能力参数）进行估计。Bock 和 Aitkin 提出了引入 E-M 算法，用边际极大似然估计方法寻找项目参数的一致估计。实际应用中，我们可以直接利用已有的参数估计软件，如 LOGIST、BILOG 和 MULTILOG 等。

三、结果与分析

1.试题信息反馈

本研究中，选择题主要采用经典测量理论的折半信度［斯皮尔曼-布朗（Spearman-Brown）公式、卢隆（Rulon）公式、弗拉纳根（Flanagan）公式］和库德-理查逊（Kuder-Richardson）信度（K-R20、K-R21 公式）进行分析。表 1 表明，每种方法计算的选择题信度都略低于 0.60，表明测试信度偏低，误差较大。提高选择题测验信度的途径主要有以下三个：一是适当增加试题量，比如从 10 道选择题增加到 12 道；二是提高命题质量，试题难度中等，区分度要大，例如替换难度过低且区分度较低的试题；三是试题编排由易到难，以消除学生的考试焦虑情绪。

表 1　选择题的折半信度和库德 - 查理逊信度

折半信度			库德-查理逊信度	
Spearman-Brown	Rulon	Flanagan	K-R20	K-R21
0.554	0.552	0.552	0.550	0.476

填空题和解答题为非 0、1 记分的项目，本研究采用克龙巴赫 α 系数进行统计，得出 α=0.740>0.50，表明填空题、解答题的信度尚可，测试真实可靠。

测验的效度则由学科专家通过逻辑分析法进行分析。根据测量的目标与内容的双向细分表，我们了解到，测验所设计的试题覆盖了所要测内容的主要方面，考查目标清晰明确、题型和分数结构合理恰当，总体符合考试大纲和教学要求。

本研究利用 CTT 和 IRT 来评价试卷的难度和区分度。通过表 2 的数据，我们可以了解到试题的难度、区分度大体合理，但仍有某些试题需要斟酌其适合性。总体说来，大部分选择题的难度偏低，而最后一道解答题的难度系数过大。试题的区分度方面，根据两种测量理论得出的结果都比较一致，但最后一道题存在着很大的差异（这在下文的试题分析中将得出结论）。选择题的猜测度方面，IRT 给出的猜测度参数 C 均较小，说明试题较真实地反映了被试

表 2　经典测量理论 (CTT) 与项目反应理论 (IRT) 测量结果的比较

	CTT		IRT						
项目	区分度	难度	a	b_1	b_2	b_3	b_4	I_{ideal}	I_{itrue}
1	0.523	0.356	0.42	–0.64				0.181	0.086
2	0.484	0.719	0.58	1.35				0.181	0.094
3	0.650	0.376	0.88	–0.39				0.181	0.290
4	0.459	0.190	0.84	–1.21				0.181	0.181
5	0.561	0.346	0.62	–0.58				0.181	0.188
6	0.457	0.262	0.56	–1.10				0.181	0.141
7	0.409	0.283	0.26	–1.80				0.181	0.033
8	0.461	0.219	0.73	–1.15				0.181	0.174
9	0.491	0.525	0.35	0.57				0.181	0.062
10	0.307	0.144	0.62	–1.86				0.181	0.079
11~15	0.432	0.453	1.31	–1.81	–0.65	0.77	3.08	0.906	0.514
16	0.634	0.413	1.58	–1.00	–0.07	0.75		0.435	0.770
17	0.403	0.405	1.20	–1.92	–0.36	1.16		0.435	0.427
18	0.353	0.705	1.35	–0.59	1.89	2.89		0.435	0.447
19	0.456	0.725	1.06	0.46	1.90	2.45		0.435	0.278
20	0.526	0.514	1.43	–1.10	–0.09	1.45		0.471	0.612
21	0.149	0.943	1.60	2.31	2.53	3.70		0.507	0.060

注：I_{ideal} 表示应提供的信息量=分数比例 * 总信息量 (划界点 θ=0)，I_{itrue} 表示实际提供的信息量；填空题和解答题的项目反应理论模型采用赛姆吉玛 (Samejima，1969) 的等级反应模型 (Grade Response Model)。

的能力。解答题方面，除了最后一题，不同分数组的难度跨度也基本均衡。在信息提供量方面，大多数选择题实际提供的信息量略低于应提供的信息量；解答题的性能相对要好些，但 19 题 (应用题) 和 21 题 (综合题) 实际提供的信息量也没有达到要求。两种测量方法都反映了选择题第 2、9 题，解答题第 18、19、21 题较难；而选择题第 3 题和解答题第 16、20 题的难度比较合理，区分度也较高。

下面对第 19 题实际提供的信息量远低于应提供的信息量加以分析。

解答题 19：某电脑公司今年 1 月、2 月、3 月生产的手提电脑的数量分别为 1 万台、1.2 万台、1.3 万台，为了估测以后各月的产量，以这三个月的参量为依据，用一个函数模拟产品的月产量 y 与月份 x 的关系。根据经验，模拟函数选用如下两个：$y=ab^x+c$，$y=-\frac{a}{x}+bx+c$ (a、b、c 为常数)．结果 4 月份、5 月份的产量分别为 1.37 万台、1.41 万台．根据上述数据，测算选用哪个函数作为模拟函数较好？并求出此函数的表达式．

分析试题本身，可以看出该题难易程度适宜，主要考查学生对函数知识点的掌握程度和在真实情境中的应用能力，体现了学生的阅读、数学和科学素养，对综合能力的考查更具真实、有效性，应该是一道质量较高的试题。但实际测量结果反映不佳，究其原因，主要在于很多数学教师在日常教学中侧重于理论知识的讲授，而忽略了对学生的应用能力的培养，结果使得学生在“题海战术”中游刃有余，一旦碰上灵活多变的实际问题却无法应对了。所以，教师在今后的教学中应该让学生多接触些与课本知识密切联系的源于生活的应用题，培养他们综合运用数学知识和方法分析解决实际问题的能力。

2.CTT 与 IRT 的实证比较

(1) IRT 更容易观测各个项目的特征属性

由表 2 可知，IRT 的测量结果和 CTT 的测量结果大体相似，但在某些项目上存在差异。为更好地观测二者之间的异同点，我们将 CTT 和 IRT 的项目参数以图表的形式进行比较，如图 3 和图 4 所示。

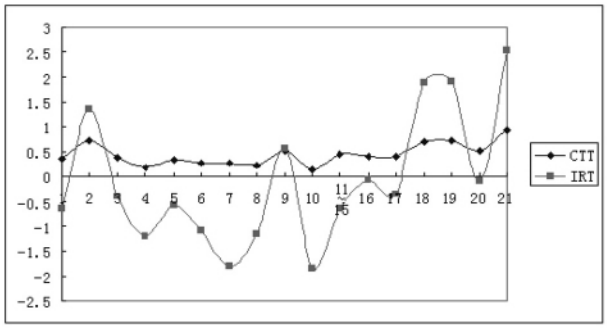

图 3　CTT 和 IRT 的难度指标比较

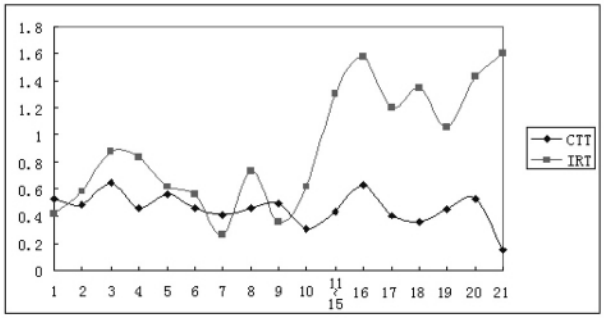

图 4　CTT 和 IRT 的区分度指标比较

通过观察图 3 和图 4 的曲线变化，我们可以发现，CTT 和 IRT 的曲线走势 (即高低变化) 大致相近，但 IRT 的变化更加鲜明一些、敏感一些，更容易观

测各个项目的特征属性。例如，IRT 在解答题上表现出来的项目区分度远大于选择题，这在实际中是比较合理的，而 CTT 则没有反映出该特点。从这一点来看，IRT 明显优于 CTT。

⑵ IRT 不依赖于样本，具有参数不变性

从图 4 可以看出，两种测量方法在第 21 题上的分析结果存在很大的差异，CTT 认为该题区分度小，IRT 则表明该题区分度很高。

解答题 21：已知函数 $f(x)=\log_2(x+1)$，当点 (x,y) 在函数 $y=f(x)$ 的图像上运动时，点 $(\frac{x}{3},\frac{y}{2})$ 在函数 $y=g(x)$ $(x>-\frac{1}{3})$ 的图像上运动.

⑴ 求 $y=g(x)$ 函数的解析式；

⑵ 求 $F(x)=f(x)-g(x)$ 函数的零点；

⑶ 函数 $F(x)$ 在 $x\in(0,1)$ 上是否有最大值、最小值；若有，求出最大值、最小值；若没有请说明理由。

经典测量理论认为 21 题的区分度很小。参看该题被试的得分分布（如图 5 所示），几乎所有被试都集中在低分段（0~3 分），仅有少数被试得 9 分。这表明，该题在考查学生的能力差别方面没有什么贡献，故区分度低。

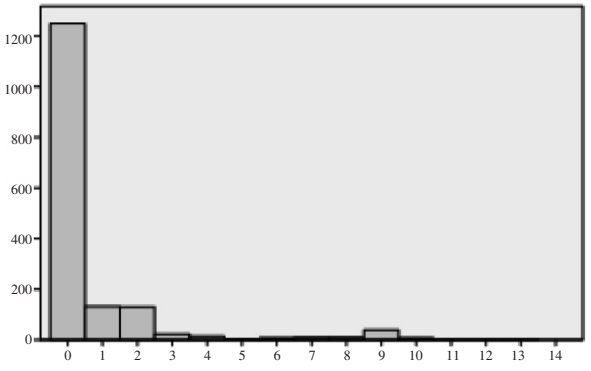

图 5　解答题 21 题的得分频数分布柱状图

但从试题本身分析，这是一道函数综合应用题，对函数的概念、性质、图像等内容掌握透彻的学生，应该可解答出大部分的试题，而没深入理解上述内容的被试，则可能无法做出正确的回应。理论上，该题能很好地鉴别学生的能力高低，这与项目反应理论的结论是吻合的，即难度大，区分度也大。

联想到很多学生在考试结束后反映试卷题目多，难度大，试题都没答完，有些学生甚至最后一道题（21 题）都没来得及看，更别提完成。从中我们不难理解该题的区分度很低的原因。因为 CTT 对被试样本有很强的依赖性，而其区分度从本质上讲是样本群体的项目分数和测验总分之间的相关关系，此时样本的抽样就会影响参数的估计，使得所估参数对测验的分析价值有限。而 IRT 的项目参数估计独立于样本，在求取项目特征曲线的参数时，用能稳定反映被试水平的潜在特质量表分数代替被试卷面分数作为回归曲线的自变量，故其难度、区分度等参数都是不变的，从而使测验结果更精确。

3.个案分析

CTT 中，被试的总分常用来作为其能力水平的参数指标。一般说来，考生的总得分与其能力水平大体一致，但也不尽相同。现实中往往存在总分相同的个体，其能力水平方面却存在着差异；总分高的被试，要视其作答情况，能力参数不一定高。IRT 将能力参数和难度参数定义在同一量表上，故能直接进行比较，这给精确测量和试卷的编制带来了极大的方便。下面，笔者选取 10 个被试样本进行被试能

表 3　个案样本得分情况与能力参数

series	1	2	3	4	5	6	7	8	9	10	11~15	16	17	18	19	20	21	sum	θ
1	5	5	5	5	5	5	5	5	5	5	20	11	8	12	11	12	9	133	2.32
2	5	5	5	5	5	5	5	5	5	5	20	12	11	4	4	10	9	120	1.96
3	5	0	5	5	5	5	5	5	5	5	20	12	10	5	9	8	1	110	1.26
4	5	0	0	5	5	5	0	5	0	5	17	12	12	7	5	8	9	100	1.28
5	5	0	5	5	5	5	5	5	0	5	15	12	12	4	4	13	0	100	1.09
6	5	5	5	5	5	5	5	5	5	5	22	6	4	5	1	8	0	96	0.65
7	5	0	5	5	5	5	5	5	0	5	18	11	2	4	0	13	2	90	0.77
8	5	0	5	5	5	5	5	5	0	5	15	6	7	4	0	8	0	80	0.24
9	0	0	5	5	5	5	5	5	5	5	15	0	8	4	0	8	0	75	0.10
10	5	0	0	5	0	5	5	5	0	5	10	3	4	1	0	2	0	50	−1.15

注：sum 表示总分，θ 表示能力值。

力的个案分析。

由表 3 我们可以看出，10 位被试的能力参数水平大体和总分的高低相符合，但也不能排除某些总得分相对不高的个体可能具有更高的能力值。例如样本 4，总分为 100，能力参数 θ_4=1.28，略高于总分为 110 的样本 3（θ_3=1.26）；样本 6 的总分 96，大于样本 7 的总分 90，但能力参数 θ_6=0.65<0.77=θ_7。就样本 4 来说，总分不太高，主要原因在于选择题的答题情况不太理想，但他在区分度和难度系数较高的解答题上表现良好，最后一题（21 题）得了 9 分，这在高分组中也是比较罕见的。由此可得，了解考生的能力水平，简单地根据总分来进行判断是不够准确和公正的，还要分析其具体作答情况。另外，总分相同的被试，其能力水平不一定相同，例如样本 4 和样本 5，总分都是 100 分，但两位的能力值（θ_4=1.28，θ_5=1.09）却不相同。

四、几点建议

1.使用 IRT 中的能力参数代替分数

本次研究表明，简单将学生的总分看成能力的指标是不科学的。在很多人的意识中，分数被认为是一个学生各方面能力的综合体现。但实质上，分数并不能承载这么多的内涵。考试分数在一定程度上可以反映学生书本知识掌握的情况，但不一定能反映学生的实际能力。按照总分得到的排名也不能作为衡量学生的综合能力的唯一标准，而只能作为一个参考。因而，我们应采用一种更客观的参数（例如 IRT 中能力参数）来代替分数，以更公正地反映学生的真正水平。

2.教师应该加强培养学生对知识的应用能力

教育测量与评价不仅要关注学校的教学效果，而且要关注学生的发展。[7]《全日制普通高中数学课程标准》指出“高中数学课程应力求使学生体验数学在解决实际问题中的作用、数学与日常生活及其他学科的联系，促进学生逐步形成和发展数学应用意识，提高实践能力”。[8]所以，数学教学向生活回归，向应用贴近，发展学生的数学应用意识和能力，是教学新课程应予努力的重要方面。

3.试卷质量评价应该质性分析与量化分析相结合

试卷的质量分析不仅要对所命制试题是否符合命题规则和考核目标等方面进行质性分析，同时也需要根据考生的作答情况进行量化分析。

笔者认为，从不同的评价视角来看待评价问题，应根据评价的目的、对象，评价的条件与环境，以及评价者自身的特点，选择适当的评价方法，以获得全面、准确的信息。质性评价强调对现象的深入了解，尊重教师和学生对自己行为的解释。量化评价能对课程教育现象的因果关系作出精确分析，得出客观科学的结论。[9]所以，教育评价工作者应将质性分析和量化研究结合起来，使教育评价方法从对立走向统一与多元，从而更逼真地反映教育现象和问题的本质。

参考文献：

[1] 教育部基础教育质量监测中心.《国家中长期教育改革和发展规划纲要》关注教育质量监测[N].基础教育质量监测信息简报，2010.1.

[2] 陈玉昆.教育评价学[M].北京：人民教育出版社，1999.

[3] 郭熙汉，何穗，赵东方.教学评价与测量[M].武汉：武汉大学出版社，2008.

[4] 杜洪飞. 经典测量理论与项目反应理论的比较研究[J].社会心理科学，2006(6)：15~17.

[5] Christine DeMars. Item Response Theory [M]. London: Oxford University Press，2010.

[6] 漆书青，戴海琦，丁树良.现代教育与心理测量学原理[M].北京：高等教育出版社，2002.

[7] 黄光扬. 教育测量与评价[M]. 上海：华东师范大学出版社，2002.

[8] 中华人民共和国教育部. 全日制普通高中数学课程标准(实验)[M].北京：人民教育出版社，2003.

[9] 张丽华. 定性与定量研究在教育研究过程中的整合[J].教育科学，2008(6)：35~38.

责任编辑/王彩霞