高考数学中考试评价的研究*

——基于 CTT 与 IRT 的实证比较

闫成海[1] 杜文久[2] 宋乃庆[2] 张 健[3]

(1. 西安文理学院数学与计算机工程学院, 西安 710065;2. 西南大学数学与统计学院, 重庆 400715;
3. 重庆市教育考试院,重庆 401147)

摘 要:相关研究表明,IRT 在教育考试评价中比 CTT 具有诸多优点。本文以某地区高考数学考试数据为基础,比较 CTT 与 IRT 在项目参数、评价方式、精度估计三个方面之间的差异。研究结果证明,在 IRT 下参数更容易反映观测各个项目的特征属性,IRT 参数比 CTT 参数更具精确性,项目信息函数能更好的反映试题信息; CTT 与 IRT 的评价方式不同,IRT 下的能力分数优于 CTT 下的测验分数,更能反映学生能力水平; CTT 与 IRT 精度估计不同,IRT 测验信息函数和能力置信区间比 CTT 有更好的精度。实证展示出 IRT 在高考数学考试评价中的优越性,具有重要的价值和应用前景。

关键词:CTT; IRT; 考试评价

一、问题的提出

高考对试题命题和质量的评价至关重要。目前,试卷的制定和评价主要是基于经典测量理论(Classical Test Theory,CTT)和项目反应理论(Item Response Theory, IRT)。CTT 利用桑代克(E. L. Thorndike)"凡物之存在必有其数量"和麦柯尔(W. A. McCall)"凡有数量的东西都可以被测量"作为理论依据①。根据学生的考试分数进行直接评价,也称为真分数理论。CTT 理论试卷评价方法简单、运算方便,易于掌握,是我们目前广泛熟悉和应用的测量理论。它对试卷的评价主要是依靠试题的难度、区分度、效度和信度进行。除了难度是一个比例之外,其余三个指标都是依靠相关性概念来对试卷进行评价分析。CTT 理论依靠样本,样本不同对同一份试题的评价也就会产生差别。IRT 也称潜在特质理论,起源于 20 世纪三四十年代的心理测量研究。基于一定假设,用一个数学函数去刻画被试在项目上可观察的作答表现(得分)与其不可观察的特质水平(能力)之间的关系,利用这个函数关系,可以对被试在项目上的作答反应进行预测,同时也可以利用被试在项目上的作答反应对被试的能力进行估计。可以说,模型与假设是整个 IRT 的核心和基础。目前比较常用的数学模型是二参数逻辑斯蒂模型、三参数逻辑斯蒂模型、Rasch 模型和等级评分模型②。

IRT 已成为一种新的现代心理与教育测量理论,如 SAT、PISA 等考试,都是基于 IRT 的应用。我国现在大学英语四、六级考试也开始运用 IRT 进行等值研究③。王晓华④、沈南山⑤、赵守盈⑥等人分别就 IRT 在教育考试命题质量、学业测试、标准化考试等方面进行了研究。但是这些研究都还不涉及实际的普通高考。为此,本文以某地区高考数学数据为例,从项目参数、评价方式和试卷估计精度对 CTT 与

* 基金项目:教育部哲学社会科学研究后期资助项目"中国基础教育改革与发展研究"(项目批准号:11JHQ001),重庆市教育科学规划项目"项目反应理论在普通高考中的应用"(项目编号:2011 KS035)。

IRT 进行比较分析，以期能为 IRT 应用于高考数学考试提供一种探索性模式。

二、考试数据的结果

在这次的高考中，数学试卷包含了填空题、选择题、解答题共 3 个大题，其中填空题包含 5 个小题，选择题包含 10 个小题，解答题包含 6 个小题，共有 21 个小题。有十多万被试参加了当年的考试，数据处理采用了 IRTP 软件和 EXCEL 进行处理，结果如表 1 所示。

在用 IRT 分析测验数据时，首先需针对不同的项目选择不同的模型。填空题选用二参数逻辑斯蒂模型，选择题选用三参数逻辑斯蒂模型，并且 c 参数取为 0.25，解答题选用等级评分模型。试题解答是需要设置步骤的，并根据参考答案的给分步骤，也相应设置了节点（得分点），全卷一共有 40 个节点。在 CTT 中，对选择题和填空题的项目难度定义为被试在项目上的正确反应比例，解答题的难度定义为被试在项目上的平均分比项目总分，项目难度的取值范围在 0 ~ 1 之间，难度值越大，项目反而越简单，也就是说项目的难易程度与难度指数的大小是反序的。项目区分度则定义为被试在测验中获得的总分与项目分数之间的相关系数，由此得到的区分度也叫内部一致性系数。

表 1　IRT 与 CTT 项目指标分布图

项目	a	b	c	难度	区分度
1(1)	0.18	-9.90	0.25	0.94	0.42
2(1)	1.30	-1.22	0.25	0.89	0.52
3(1)	1.48	-1.38	0.25	0.92	0.52
4(1)	1.43	-1.11	0.25	0.89	0.54
5(1)	1.27	-1.26	0.25	0.90	0.52
6(1)	1.70	-0.82	0.25	0.86	0.58
7(1)	0.95	-0.05	0.25	0.65	0.44
8(1)	1.70	0.03	0.25	0.66	0.50
9(1)	2.66	0.77	0.25	0.38	0.40
10(1)	2.01	0.97	0.25	0.39	0.24
11(1)	1.28	-1.64	0.00	0.92	0.52
12(1)	1.75	0.03	0.00	0.54	0.62
13(1)	1.29	0.07	0.00	0.51	0.55
14(1)	1.54	0.72	0.00	0.25	0.46
15(1)	1.18	1.61	0.00	0.08	0.27
16(1)	3.51	-0.74	0.00	0.68	0.80
16(2)	2.37	-0.38	0.00	0.68	0.80
16(3)	1.88	-0.12	0.00	0.68	0.80
16(4)	1.52	0.45	0.00	0.68	0.80
17(1)	1.77	-0.68	0.00	0.74	0.74
17(2)	1.56	-0.21	0.00	0.74	0.74
17(3)	1.43	-0.07	0.00	0.74	0.74
18(1)	2.40	-0.72	0.00	0.63	0.80
18(2)	2.17	-0.53	0.00	0.63	0.80
18(3)	2.16	-0.41	0.00	0.63	0.80
18(4)	1.79	0.58	0.00	0.63	0.80
18(5)	1.46	1.05	0.00	0.63	0.80
19(1)	1.56	-1.00	0.00	0.56	0.75
19(2)	2.05	0.20	0.00	0.56	0.75
19(3)	2.44	0.52	0.00	0.56	0.75
19(4)	2.16	0.67	0.00	0.56	0.75
20(1)	1.92	-0.17	0.00	0.42	0.75
20(2)	1.82	-0.02	0.00	0.42	0.75
20(3)	2.71	0.50	0.00	0.42	0.75
20(4)	3.55	0.95	0.00	0.42	0.75
21(1)	1.92	-0.07	0.00	0.22	0.68
21(2)	1.97	0.43	0.00	0.22	0.68
21(3)	111.30	288.83	0.00	0.22	0.68
21(4)	111.14	318.49	0.00	0.22	0.68
21(5)	111.01	320.20	0.00	0.22	0.68

说明：平均分：90，标准差：31.69，信度 rx：0.84，测验标准误：12.52

(一) CTT 下的结果分析

在 CTT 下的难度与区分度参数分布如表 2。从表 2 可知,在该次考试中,信度系数为 0.84。难度指数小于 0.3 的试题有 3 题,位于 0.3 至 0.7 之间的试题有 10 题,大于 0.7 的试题有 8 题。区分度指数除了有两个题小于 0.3 以外,其余的值均大于 0.3。因此,从 CTT 的观点来看,该次考试的难度中等偏易,质量较好。

表 2 难度与区分度参数分布表

难度区、分度分布范围	0 ~ 0.30	0.30 ~ 0.70	0.7 ~ 1.00
难度题数	3	10	8
区分度题数	2	15	4
平均分	标准差	信度	测验标准误
90	31.69	0.84	12.52

(二) IRT 下的结果分析

在 IRT 下的难度与区分度参数的分布如表 3。从表 3 看到,项目难度或类别难度参数 b 在 -2 以下的有 1 个,位于 -2 ~ 2 内的项目参数或类别参数有 36 个,大于 2 的类别难度或项目难度参数有 3 个。项目或类别区分度参数 a 小于 0.5 的有 1 个,0.5 ~ 2 的项目有 24 个,2 以上的项目有 15 个。

表 3 项目难度与区分度参数分布

难度 b	-2 以下	-2 ~ 2	2 以上
题数	1	36	3
区分度 a	0.5 以下	0.5 ~ 2	2 以上
题数	1	24	15

在 IRT 中,难度参数 b 的取值范围为一切实数,一般要求 b 参数位于 -2 ~ 2 之间[⑦],b 参数过大与过小的项目都不利于对被试的能力参数进行有效估计。在本次考试中,有 36 个项目或类别 b 参数位于 -2 ~ 2 之间,因此从 IRT 角度看,这 36 个项目(或类别)的 b 参数是合适的,但是项目 21 有 3 个类别 b 参数都大于 200。从 IRT 角度看,这样的试题是过难的。因为无论是高能力的被试或者是低能力的被试都无法对这样的试题做出正确反应,因此这样的试题不能对被试的能力进行有效的鉴别。另外有一道选择题的难度参数为 -9.9,它意味着几乎所有的被试都能对该试题做出正确反应,这样的试题仍然不能对被试的能力进行有效鉴别。在 IRT 中,a 参数在理论上可以取一切正实数,但是为了对试题(类别)参数及被试的能力参数进行有效估计,一般要求 a 参数位于 0.5 ~ 2 之间[⑧],过大或者过小的 a 参数都会对参数的估计精度带来不利影响。然而在表 3 中看到,有一个试题的 a 参数小于 0.5,有 15 个试题或者类别 a 参数大于 2,因此从 IRT 角度看,这些试题的 a 参数是不理想的。特别是第 21 题有 3 个类别 a 参数的估计值大于 100。第 1 题的 a 参数只有 0.18,这样的试题对被试的能力估计几乎没有任何贡献。当然这样的结果可能与这套试题是基于 CTT 制定有关。

三、CTT 与 IRT 项目参数的比较

(一) CTT 与 IRT 项目难度参数的比较

从表 1 中可知,当 CTT 中项目难度值相同时,它所对应的 IRT 中的难度参数值有些差别不大,如第 2 题和第 4 题,这是两个选择题,在各节点的难度参数都为 0.89。它各节点所对应的 IRI 难度参数分别为 -1.22 和 -1.11。有些题目差别就大一些,如 20 题第 2 节点和第 4 个节点,CTT 难度参数为0.42,IRT 难度参数却分别为 -0.02 和 0.95。这就是说,对于相同的试卷,CTT 项目难度参数相同时它在 IRT 中的难度参数并非一致。

CTT 与 IRT 难度参数比较如图 1 所示，横坐标是试题数目，3 表示第 3 题，16.4 表示第 16 题的第 4 个节点，纵坐标表示取值。由于 IRT 里面的 21 题第 3 步以后的题目难度区分度值太大，故在对比图里面没有画出。

从图 1 中可以看出，CTT 的难度参数和 IRT 的难度参数大体相似，但在某些项目上存在差异。可以发现，CTT 和 IRT 的项目难度曲线走势（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。⑨

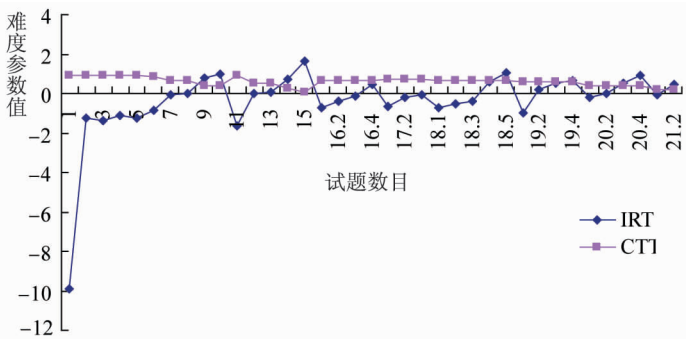

图 1　CTT 与 IRT 难度参数对比图

（二）CTT 与 IRT 项目区分度参数的比较

从表 1 可以看出，当 CTT 中项目区分度参数值相同时，它所对应的 IRT 中项目区分度参数值差别不大，如第 2 题和第 3 题。这是两个选择题，在 CTT 下的区分度参数都为 0.52，在 IRT 下的区分度参数分别为 1.30 和 1.48。有些题目差别就大一些，如第 20 题第 3、4 节点，CTT 区分度参数为 0.75，IRT 却分别为 2.71 和 3.55。在 CTT 下区分度参数值为 0.75，这是一个尚可的值，在 IRT 下的值为 2.71 和 3.55，却是一个较差的值。这就说，对于相同的试卷，CTT 项目区分度参数相同时它在 IRT 中的区分度参数并非一致。

CTT 与 IRT 区分度参数的比较如图 2 所示。从图 2 可以看出，区分度参数具有难度参数同样的特征，IRT 区分度参数更容易观测各个项目的特征属性。

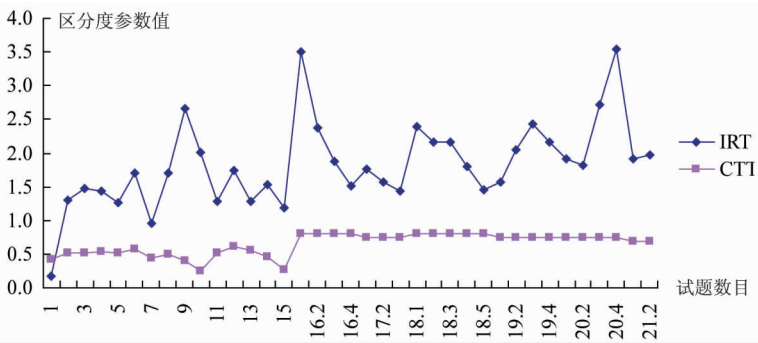

图 2　CTT 与 IRT 区分度参数对比图

（三）CTT 与 IRT 中难度与区分度参数的比较

当 CTT 中区分度与难度参数一致时，它所对应的 IRT 中区分度与难度参数值差别不大，如第 3 题和第 11 题，在 CTT 中区分度与难度参数值一致，分别为 0.52 和 0.92，在 IRT 中所对应的区分度与难度参数却是不同的，第 3 题区分度和难度参数分别为 1.48 和 −1.38，第 11 题区分度与难度参数分别为 1.28 和 −1.64。有些题差别就大一些，如 20 题第 2 和第 3 节点，在 CTT 中区分度与难度参数为 0.75 和 0.42，在 IRT 中区分度与难度参数却分别为 1.82、−0.02 和 2.71、0.50。

综上可知,CTT 参数在反映试题的难度和区分能力上有些粗糙,IRT 参数比 CTT 参数更精确的反映试题参数问题。

(四)项目信息函数

在 CTT 中对试题的评价主要是基于难度和区分度。IRT 的试题评价不仅仅是难度和区分度这两个指标,重要的是引入项目信息函数这个概念。例如第 11 题的项目信息函数图如图 3。

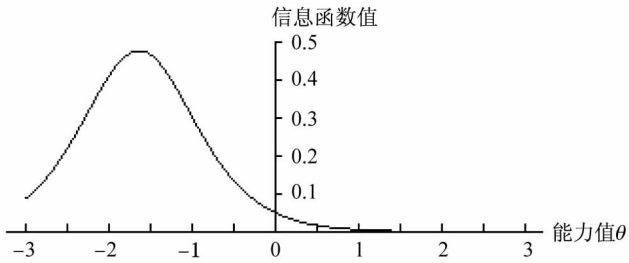

图 3 第 11 题项目信息函数

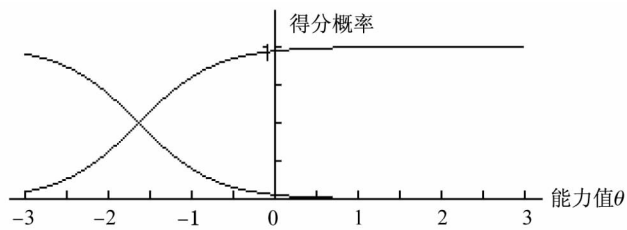

图 4 第 11 题项目特征曲线

从图 3 可知,第 11 题的项目信息函数值在 0.5 附近,它所提供的信息一般。在 $\theta = -1.6$ 时,达到峰值,对于能力 −1.6 的被试提供了最大的信息。在能力大于 和小于 的被试提供了较少的信息,这个题目适合低水平能力的被试。它的 IRT 难度与区分度参数分别为 −1.64 和 1.28,项目特征曲线如图 4,也是被试得 0 分和 1 分的概率图。IRT 对题目的评价主要是看该试题与这个能力段的被试是否匹配。在 CTT 下第 11 题的难度是 0.92,区分度是 0.52。它的难度不好,但区分度较好。再比如,第 12 题的项目信息函数如图 5。

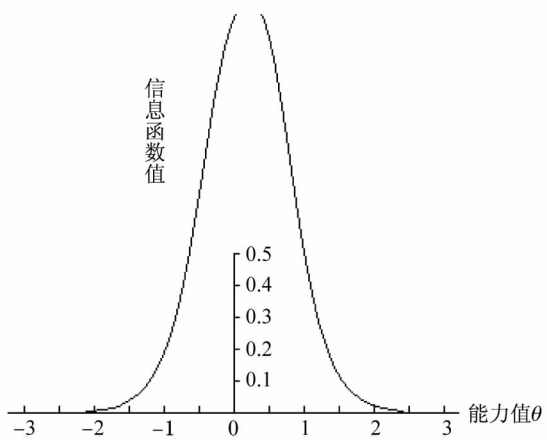

图 5 第 12 题项目信息函数

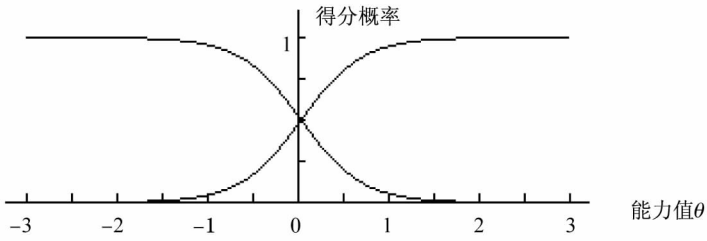

图 6　第 12 题项目特征曲线

从图 5 可知，第 12 题的项目信息函数值远远大于 0.5，它提供的项目信息很好。在（－0.5，0.5）提供了较多的信息，对在这个能力区间的被试提供了较大的信息，尤其对于能力 0.2 附近的被试提供了最大的信息量，对于能力大于 1.5 和能力小于－1.5 的被试提供的信息较差。它的 IRT 区分度与难度参数分别为 1.75 和 0.03，项目特征曲线如图 6。CTT 难度与区分度参数分别为 0.54 和 0.62，说明 CTT 下试题区分度较好。从上可知，CTT 是绝对的，IRT 对试题进行评价更精细、更客观，而且是相对的。

四、CTT 与 IRT 评价方式的比较

在 CTT 中以学生的测验分数代替学生的能力，所有被试的数学成绩分布如图 7 所示。

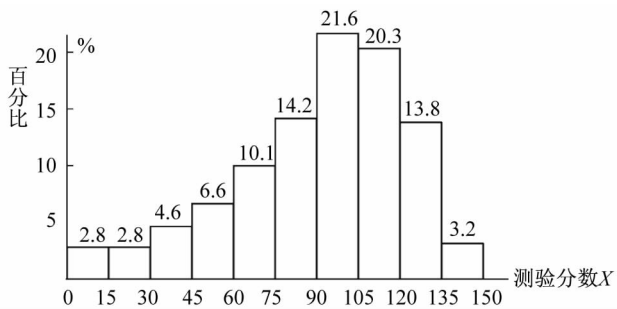

图 7　测验分数分布

从图 7 可知，被试的测验分数分布呈现明显偏态分布，其峰值位于 90 分至 105 分之间，高分数段的被试所占比例较多，低分数段被试所占比较小。这说明当年高考数学试题偏易，这与难度指数的分布情况是一致的。

在 IRT 中主要用能力参数描述被试的学业成就，由于人们对能力参数不习惯，为此可以将能力参数转换为人们熟悉的“分数”。设

$$X=\begin{cases}0, & \theta\leqslant -2.5\\ 30(\theta+2.5), & -2.5<\theta<2.5\\ 150, & \theta\geqslant 2.5\end{cases}$$

通过上述转换，X 的取值范围为 0 ~ 150，与测验分数的取值范围一致。由于 X 是由能力参数转换得到的，因此我们称 X 为能力分数。能力分数估计量是相合估计。就是说，如果某一被试的能力分数真值为 X_0，$\hat{X}$ 是被试的能力分数估计值，那么，当试题样本容量 $n\to\infty$ 时，$\hat{X}$ 将依概率收敛于真值 X_0。测验分数不具有这样的性质。因为在 CTT 中，总分是固定的，当试题增加时，每一题的得分就要重新划定，这时测验分数的意义已经不是原来意义上的分数了。只有在同一个测验中重复做无穷多次，被试的测验分数才是相合的。然而在实践中，这是很难做到的。由于能力参数具有不变性，因此由能力参数转换而得到的能力分数也同样具有不变性这一性质。换句话说，被试在测验中即可参加 A 卷测验，

也可参加 B 卷测验,除去抽样误差外,将获得相同的能力估计。

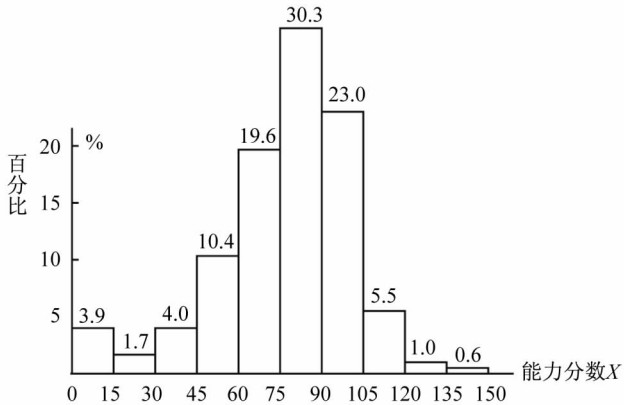

图 8 能力分数分布

所有被试的数学能力分数分布如图 8 所示。从图 8 可以看出,被试的能力分数分布与测验分数分布存在较大差异,能力分数分布呈现明显的正态分布特征,其峰指出现在 75 ~ 90 分之间,占被试总数的 30.3% 。能力分数位于 105 ~ 120 分与 120 ~ 135 分之间的被试分别占总数的 5.5% 和 1% ,测验分数占比分别为 20.3% 和 13.8% 的比例均有较大幅度降低。位于 135 ~ 150 分之间的被试也由 3.2% 降低到 0.6% 。这表明在 IRT 框架下,去掉了一些虚假的高分,使分数的分布更趋于合理。

五、CTT 与 IRT 估计精度的比较

在 CTT 中,对测验精度主要用信度和测验标准误来进行刻画。该次数学考试的信度系数为 0.84,测验的标准误 12.52,学生测验分数与真实分数之间的平均误差是 12.52。信度是一个笼统的、粗略的指标,它只是大致的描述了被试的测验分数与真实分数之间的平均误差。

在 IRT 中,刻画试卷信度是利用测验信息函数这个概念,整体评价。试卷的测验信息函数如图 9 所示。

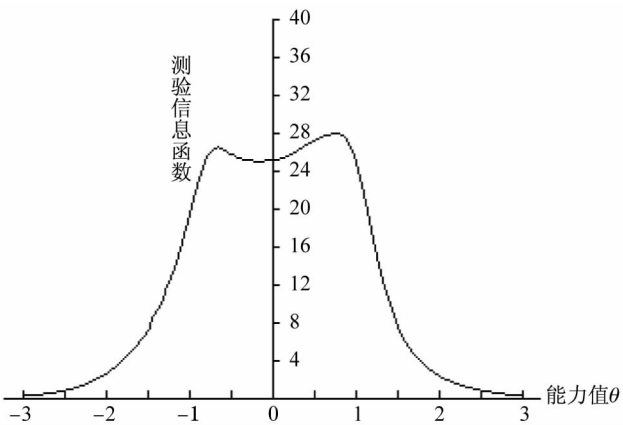

图 9 测验信息函数

从图 9 中可以看出,在区间(-1,1)内,测验信息函数值均大于 25,该测验提供了较大的信息量,而在这之外则提供了较少的信息量,说明这是一次不错的测验。从图 9 可以看出,该图呈双峰型,在能力值 -0.8 和 0.9 附近,该项目的信息量分别达到了两个不同的峰值。而在(-0.5,0.4)之间存在一个凹区间,因此在这个区间提供的信息量较少。

在 IRT 中,刻画测验误差的方法则是置信区间, IRT 能力分数估计 95% 的置信区间为

$(\hat{X} - 1.96 \cdot \frac{30}{\sqrt{I(\theta)}}, \hat{X} + 1.96 \cdot \frac{30}{\sqrt{I(\theta)}})$，其中 $I(\theta)$ 是测验信息函数[10]。

该次测验能力分数估计值95%的置信区间如图10所示，其中，横坐标为能力分数的估计值，纵坐标表示能力分数真值，下曲线表示置信区间的左端点曲线，上曲线表示置信区间的右端点曲线。比如假设某被试的能力分数估计值为75分，那么在95%的意义下，该被试的真实分数约位于64～85分之间。

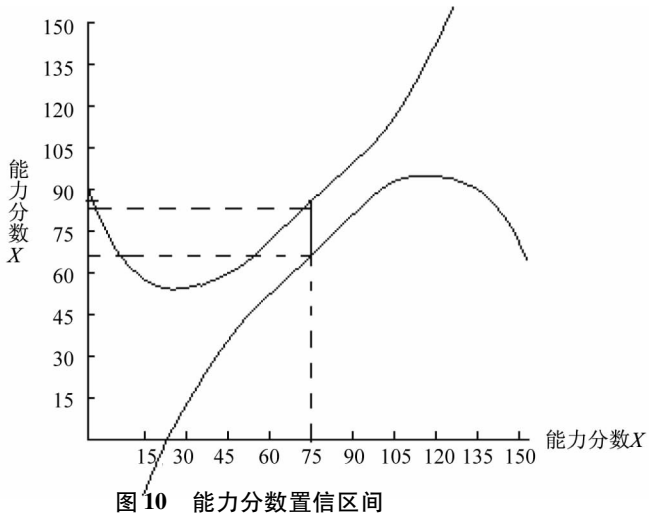

图10　能力分数置信区间

从图10看到，在该次考试中，对能力分数位于45～105分的被试的估计精度较高，其估计误差略为11分左右，而对能力分数估计值位于45～105分以外的被试，其估计误差较大。特别是对能力分数估计值大于130和小于30分的被试，其估计误差大于30分，这样大的估计误差实际上已经没有多大的意义。这一结果说明，在同一次测验中，对于不同能力的被试，其能力分数的估计误差也不相同。

六、讨论

通过上面的数据分析看到，IRT克服了CTT中的许多缺点，主要表现在以下4个方面：

1. 在IRT框架下，它的项目参数比CTT框架下的参数值更具有精确性。

2. 能力分数分布优于测验分数分布。这表明在IRT框架下，去掉了一些虚假的高分，使分数的分布更趋于合理。

3. IRT比CTT有更好的估计精度。IRT定义了CTT中没有的项目信息函数和测验信息函数，它是一个具体地、动态地刻画项目和测验性能的综合指标。它指出了每个项目在不同能力水平处提供的信息量的大小，IRT抛开了平行形式的信度观念，直接面向测量标准误，用信息函数来计算估计精度。

4. IRT提出试题编制信息量最大原则。IRT提出了测验编制的指导原则，以项目难度与考生能力水平匹配的原则，即信息量最大原则，在实际编制测验时以信息量为指导的原则。

然而，也不能忽视IRT存在的一些不足。目前，在IRT下能力是基于单维性假设。实际上，被试的能力不止一种，IRT也从单维研究走向多维[11]，多维能力参数估计还处于研究之中。应用也需要一定的软件支撑。对项目研究也有一些偏差，个别项目上CTT参数较好，IRT参数值却比较差，这些都还需要继续研究。

总之，虽然IRT目前存在着一些缺陷，但是在教育考试中尤其是高考数学考试中使用IRT进行测验的编制、报告被试的能力水平和项目性能的解释在理论上比CTT更严格、更完备，在实践中也更有效、更公平。

注 释:

①朱德全、宋乃庆:《教育统计与测评技术》,重庆:西南师大出版社,2008 年第 67 页。

②⑩杜文久:《高等项目反应理论》, 重庆:西南师大出版社,2007 年第 71 - 88、153 - 156 页。

③陈谨、何静等:《英语标准化考试评价中 IRT 与 CTT 的比较研究》,《数学的实践与认识》2011 年第 20 期。

④王晓华、文剑冰:《项目反应理论在教育考试命题质量评价中的应用》,《教育科学》2010 年第 3 期。

⑤沈南山:《基于 IRT 模型的数学学业成就水平测试分析》,《安徽师范大学学报》(社科版)2012 年第 1 期。

⑥赵守盈、石艳梅等:《项目反应理论在大规模选拔性考试试题质量评价中的应用》,《教育学报》2013 年第 1 期。

⑦ Lord, F. M. Applications of item response theory to practical testing problems. Hillsdale, NJ: Lawrence Erlbaum Associates,1980.

⑧ Hambleton, R. K. Swaminathan, H. Item Response Theory: Principles and Applications. Kluwer - Nijhoff Publishing, 1985.

⑨何穗、吴慧萍:《基于教育测量理论的中学数学试卷质量评价研究》,《教育测量与评价》(理论版)2012 年第 8 期。

⑪丁树良、罗芬等:《项目反应理论新进展专题研究》,北京:北京师范大学出版社,2012 年第 109 页。

(责任编辑 陈振华)

Evaluation of Examination in Math in College Entrance Examination: A Empirical Comparative Study on CTT and IRT

YAN Chenghai[1] DU Wenjiu[2] SONG Naiqing[2] ZHANG Jian[3]

(1. School of Mathematics and Computer engineering, Xi'an University, Xi'an 710065, China;
2. School of Mathematics and Statistics, Southwest University, Chongqing 400715, China
3. School of Chongqing Educational Examination, 401147, China)

Abstract: The previous research shows that ITR has more merits than CTT in the evaluation of examination. Based on the data collected in math subject in National College Entrance Examination of one district, the paper compares the differences between CTT and IRT on project parameters, ways of evaluation and accuracy assessment. The research shows that the characteristics of each project properties are more easily reflected under the IRT parameters, and the IRT parameters are more accurate than the CTT. The function of the project information can reflect the information of the tests better. IRT and CTT have different evaluation ways and the scores under the evaluation of IRT, which represents students´abilities, are higher than those of CTT. CTT is different from IRT in precision estimation, but the test information function and confidence interval of IRT are more accurate than that of CTT. The Empirical study shows the advantages of IRT in evaluating the math subject in College Entrance Examination, which is valuable and has potential applications.

Keywords: CTT; IRT; Evaluation of Examination