DOI:10.19789/j.1004-9398.2018.04.011

整合 CTT 和 IRT 技术 提高监测工具质量

——以小学科学试卷为例

周丐晓[1,2] 刘恩山[2*]

(1. 温州大学生命与环境科学学院,浙江 温州 325035; 2. 北京师范大学生命科学学院,北京 100875)

摘 要

试题是科学学业质量监测中常用的测试工具,也是开展教育实证研究的重要工具,工具的质量直接决定学业测评和教育研究的科学性及可靠性. 基于经典测验理论和项目反应理论系统阐述试卷质量分析的常规过程及技术方法,并对工具质量评估相关参数的内涵进行深入解读和分析,以帮助学业测评开发人员以及教育研究者系统理解和掌握试卷的质量分析方法和技术,从而有效提高学业测评和教育研究工具的质量,为进一步提高学业质量测评和教育研究的科学性及有效性奠定基础.

关键词: 经典测验理论,项目反应理论,试卷,质量评估.
中图分类号: G424. 74

当前学业质量监测作为教育质量提升的重要途径日益受到各国的重视,学业质量监测的结果可以为国家和地方决策、教育教学改进提供重要信息和客观反馈,有益于国家教育改革的发展. 作为学业质量监测的核心话题,工具质量的优劣直接影响了测试结果的信效度. 试题是监测工具中最为常见和重要的工具,如何对试卷的质量进行评估和分析,哪些技术可以帮助获取试卷质量相关的各种信息等均是试卷质量评估中亟待探讨的议题. 当前主要依据经典测验理论(classic test theory,CTT)或项目反应理论(item response theory,IRT)分析工具的质量. CTT 的优势在于便于获取被试和工具的整体测试信息,参数和估算方法易于理解和掌握,但相比 IRT 更易受样本影响,而 IRT 中所用的项目参数不受样本影响,且基于 IRT 的测量能够将误差具体到个人,更为精确也更能反映客观的被试情况. 因此结合 IRT 和 CTT 综合分析试题,对试题个体及工具整体的质量进行分析,可有效提高工具的科学性.

1 科学学科测试工具质量评估的测评理论基础

在测量理论的发展过程中,CTT 和 IRT 在心理学与教育测量方面发挥了重要作用. 当前在学科测试工具质量评估方面,主流方向是结合 IRT 和 CTT 综合分析工具,通过难度、区分度、信效度等分析,对每个试题及工具整体的质量进行分析,从而提高工具的科学性和有效性.

1.1 经典测验理论(classic test theory,CTT)

相比 IRT,CTT 出现较早,理论研究较多、体系较为完善,其基本思想是: $X = T + e$(X 指观察值,T 为真实值,e 为误差),即观察值为真实值与误差之和,由此可以发现 CTT 所涉及的数学模型相对简单,参数和估算方法易于理解和掌握,对研究者统计学原理知识的掌握程度要求不高,便于操作使用,因此在过去的几十年应用广泛. 随着测量理论的不断发展,CTT 渐渐展现出其理论和方法体系的弱点,包括: 项目难度与被试能力互相依赖,这就导致项目与被试能力之间无法进行等值化比较,各参数受样本质量的影响; 项目之间平行、不区分问题重要性; 相比 IRT 的测量结果,CTT 的测量结果信度较低,这是因为 CTT 的信度针对的是全体被试者并且假设所有被试的测量标准误差均相等,统计量(难度、区分度、误差等)是笼统的全组被试的平均值,因此 CTT 的信度仅能代表平均测量精确度,然而这一假设与真实情况不符,在真实情境下不同水平被试的测量标准误差往往不相等; 最后,CTT 缺乏预测推理能力等.

收稿日期: 2017-09-26
* 通信作者

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

1.2　项目反应理论(item response theory,IRT)

针对上述 CTT 理论的不足,项目反应理论应运而生,目前该理论被广泛应用于学业评估等多方面的研究中. IRT 理论是一种现代测验理论,也被称作项目特征曲线理论(item characteristic curse theory) 或者潜在特质理论(latent trait theory) ,它克服了经典测验理论(CTT) 的缺点. CTT 是通过被试对所有项目的反应总和(即测验的总分) 来预测被试的某一潜在心理特质,它认为被试对单个项目的反应与其测量心理特质之间没有联系. 而 IRT 的基本思想是: 被试能力与被试在某一具体项目上的反应存在某种函数关系,以 θ 表示被试者的某一潜在心理特质,用 $P(\theta)$ 表示该被试在项目 i 上作出正确反应的概率,那么根据 IRT 理论即可推断出 θ 与 $P(\theta)$ 间的函数关系,并在此基础上做出进一步的预测,即 IRT 是一种探讨被试对项目的反应与其潜在特质之间关系的概率性方法,它根据一定的数学模型,用项目参数去估计被试的潜在心理特质. 相比 CTT 易受样本影响的特点,IRT 中所用的项目参数(如题目难度、区分度等) 不受样本影响,被试能力与难度参数相互独立,这些参数的获得不会因被试样本的变化而变化,同时对被试能力的估计不会因为试题的不同而不同[1]. 其次,它将定序测量转化为等距测量,将项目难度与被试放在同一量尺上进行测量,便于比较操作. 最后,基于 IRT 的测量能够将误差具体到个人,更为精确也更能反映客观的被试情况.

2　基于经典测验理论的科学试题质量评估方法及关键技术

基于经典测验理论的试题质量检验参数主要包括: 难度、区分度、选项分析等.

2.1　试题总体的描述性分析

研究者获得试题作答后,首先需对试题整体概况和分布特征进行描述性分析. 主要包括: 被试总分、平均分、众数、最高(低) 分、峰度与偏度、选项分析. 总分、平均分、众数和最高(低) 分可描述被试的整体得分和分数特征情况,而峰度与偏度都用来描述被试总分的分布形态.

其中峰度描述的是数据围绕平均分分布的密集情况,峰度的绝对值数值越大表示被试分数分布形态的陡缓程度与正态分布的差异程度越大[2]. 峰度可通过观察总体中所有取值分布形态的陡缓程度获得,峰度为 0 表示被试总分分布曲线与正态分布的陡缓程度相同; 峰度大于 0 表示被试总分分布曲线与正态分布相比更为陡峭,为尖顶峰,表明被试数据分布较为密集地分布在本卷平均分周围; 峰度小于 0 表示被试总分分布曲线与正态分布相比较为平坦,为平顶峰,表明被试数据分布较为分散. 偏度描述数据分布的对称性和偏斜程度,通常以均值与众数(中位数) 差值和标准差的比值衡量数据的偏斜程度,偏度的绝对值数值越大表示其分布形态的偏斜程度越大,描述偏度的统计量同样需要与正态分布相比较. 偏度为 0 表示其数据分布形态与正态分布的偏斜程度相同; 偏度大于 0 表示其数据分布形态与正态分布相比为右偏态(正偏态) ,即数据右测尾巴更长,数据右端有较多的极端值,总体均值在众数的右边,表明被试总分偏低,试题难度偏大; 偏度小于 0 表示其数据分布形态与正态分布相比为左偏态(负偏态) ,即数据左测长尾更长,总体均值在众数的左边,表明被试总分较高,试题难度较易. 试题开发者和教育研究者可根据峰度和偏度综合分析试题相对特定被试的难易程度,以帮助进一步修订试题,开发出与被试水平适切度较高的试题.

此外,试题中的选择题还可进一步分析被试的选项作答情况. 选项包括正确选项和若干错误选项. 错误选项应具有一定干扰和诱答作用,能够迷惑部分学习水平稍弱的学生选择干扰项,从而探查出学生学习困难及薄弱点,同时也能提高试题的区分度. 诱答选项在设计上宜以学生看似熟悉、但并不正确的陈述或者常犯的错误为依据,使用专业的语言来撰写诱答选项. 另一方面,通常认为若某一选项没有被试选择或者选择率低于 5% ,则说明该选项的诱答作用太弱,不具备干扰功能,可考虑修改或者删除.

2.2　难　度

经典测试理论中试题的难度指的是试题的难易程度,一般用得分率来表示,对于选择题,难度计算的方法是“P = 答对此题的人数/作答的总人数”; 对于非选择题,难度计算的方法是“P = 所有人在此题上得分的均值/此题的分值”. 难度取值范围在 0 ~ 1 之间,难度数值越小表示试题难度越大,学生在该试题上的得分越低,反之难度越小. 通常认为 0.6 ~ 0.75 为中等难度,大于 0.75 表示试题得分率高较为容易,小于 0.6 则表示试题较难. 考生对试题内容的熟悉程度、考察行为目标的层次、试题情境材料的信息量、试题的形式以及测试时间均会在一定程度上影响试题的难度值.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

2.3 区分度

经典测试理论对试题的区分度并没有精确的定义. 一般可认为,试题的区分度即指试题有效区分学生某种心理特征不同水平的程度[3]. 即指试题在区分高水平和低水平学生的程度和能力,取值范围在 -1 ~1 之间,区分度越高,表明该题在区分不同水平的被试上具有更大价值. 通常认为区分度大于 0.4 即为区分度较好.

区分度的计算有多种方法,判断区分度的指标主要包括鉴别指数和相关系数. 鉴别指数的计算选取样本中总分最高的前 27% 作为高分组,同样选取样本中总分最低的后 27% 作为低分组,由高分组在此题上的得分率减去低分组在此题上的得分率,从而得到鉴别指数. 而当前在大规模教育测评通常用试题单题与同一量表总分的相关系数计算各试题的区分度,常用的表示区分度的相关系数有 3 种: 肯德尔相关非参数(Kendall´ tau-b correlation),皮尔逊积差相关(Pearson's correlation)和斯皮尔曼等级相关(Spearman correlation). 由于采用相关系数计算区分度的方法更为普遍,以下重点阐述采用相关系数衡量区分度,表 1 具体呈现了这三种相关分析的适用情境及条件.

表 1 三类相关系数的内涵

	分布条件	适用变量	适用题型
皮尔逊积差相关	自变量等间距或等比例的正态分布	一个双歧变量,另一个连续变量 两个连续变量	对错计分的客观题
斯皮尔曼等级相关	自变量顺序等距或等比例的非正态分布或不明分布	两个变量至少有一个是顺序变量(自变量顺序定距或定比)	多等级分步计分的主观题
肯德尔相关非参数	自变量可以是顺序、不等间距或不等比例的非正态分布或分布不明	两个变量均为顺序变量(自变量顺序不等间距或不等比例)	评分变量顺序不等间距或不等比例的态度量表(如满意度)等

由表 1,实际选择哪种相关系数主要参考 3 个方面的信息: 每个变量的类型,连续变量、双歧变量或者顺序变量; 潜在的分布性质,正态分布还是非正态分布; 变量分布特征,线性的还是非线性的. 试题的区分度通常采用皮尔逊积差相关和斯皮尔曼等级相关计算.

在试题分析时,需要特别注意的是若区分度小于 0,可能的原因有: 试题的答案有误; 某一个选项的干扰性过强; 此题过难,被试胡乱作答. 要了解具体原因,还可同时参考试题难度值,若试题难度值显示通过率较低,则可判断此题的区分度小于 0 是由于试题过难导致被试胡乱作答,以此为根据可进行试题的修改.

2.4 信 度

教育研究中采集的数据是否可信,测量的结果是否有效是决定整个教育研究质量的关键要素,因此测量工具的信度和效度检验是工具开发中最为重要的分析. 信度分析是指研究所进行的测量和获得的结果一致性和稳定性程度[4],测量结果中由于随机误差因素所带来的方差变异大小,信度是效度的基础.

信度包括重测信度、复本信度、折半信度、评分者信度和同质性信度等多种类型,重测信度和复本信度旨在验证多次测量之间的稳定性和可靠性,同质性信度则是度量工具量表中的各个题目是都探查的是同一概念或结构的程度. 其中在工具研发和测量中使用较多的包括评分者一致性信度中的 Kappa 一致性系数和同质性信度中的科隆巴赫系数(Cronbach's Alpha).

在测试中,对同一组试题不同评分者的评判可能存在一定差异性,因此需要对若干个评分者的评分结果一致性进行信度分析,常用 Kappa 一致性分析. Kappa 一致性系数常被用来表示各评分者之间评分结果的一致性,它适用于类别或者名义变量资料需要做归类或者评定时评分者间归类评定为一致的程度判定指标,同时也被视为一种评分者信度指标[5]. 通常认为 Kappa 值大于 0.75 表示评价者之间的一致性较好.

Cronbach's Alpha 是测验一组同质测验总和的信度,其本质是测验量表工具中所有试题的一致性程度. 若所有试题均反映了相同的构造(如能力、态度、知识等),那么各个试题之间应该具有相关存在,若某一试题和其他试题之间没有相关存在,说明这个试题与其它试题测试的构造不同,需要考虑修改或者删除该试题. 通常 α 系数大于 0.6 认为是可接受水平,0.8 以上表示信度良好[6].

Cronbach's Alpha 是对工具中一组试题信度的分析,但无法获悉单题的信度信息,有时工具开发者还需要知道整套工具中具体哪些单题信度较低,降低整卷的信度. 可通过 Cronbach's Alpha if item

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

deleted 指标了解单题对整卷信度的影响. Cronbach's Alpha if item deleted 表示如果单独删除某一题目后,整卷的信度值. 若单题在该指标上的信度值高于整卷信度值,表示删除该题后,整卷信度提高,即该题与其它题目的同质性较低,对信度产生负面影响,可考虑修改或者删除此题,Cronbach's Alpha if item deleted 可以帮助进一步分析单题的信度情况以帮助提高整卷的信度. 综合参考 Cronbach's Alpha 和 Cronbach's Alpha if item deleted 两个信度参数可帮助试题研发者和教育研究者进一步提高试题单题以及工具整体的信度.

2.5　效　度

效度是指研究真正、正确地揭示所研究问题的本质和规律的程度,数据测量结果反映测量属性的程度或实现测量目标的程度,即效度分析就是对测量性质准确性和测量结果正确性的评价[7]. 效度有多种类型,在试题分析上,主要采用内容效度.

内容效度是指工具中的试题对测验有关知识、能力或者行为等内容范围取样选择的适当性,即考察试题是否可以代表所规定的内容. 试题的内容效度通常不用数字化的量化指标衡量,而是通过相关领域的专家根据学科内容和测试蓝图进行逻辑分析以判断工具的内容效度.

3　基于项目反应理论的科学试题质量评估方法及关键技术

由于经典测验理论的检验参数对样本有严重的依赖性,稳定性不高,因此在试题分析中还需要结合项目反应理论对试题质量进行分析. 基于项目反应理论的试题分析主要关注: 难度、项目拟合度、怀特图、信息量图等参数. 其中,项目拟合度、怀特图等是对工具效度的检验,为工具效度分析提供更为科学和客观的信息.

3.1　难　度

IRT 难度与 CTT 难度的意义相同,但是 IRT 理论所采用的数学模型与 CTT 不同,通常 IRT 难度正常的取值范围为 $-3 \sim 3$ 之间,数值越大,表示试题越难,即学生正确作答的可能性越小. IRT 难度比 CTT 难度的计算更为复杂,也相对更为精细准确. IRT 理论中,一般要求试题的难度应该介于 $-2 \sim 2$ 之间.

3.2　项目拟合度

项目拟合度分析是指收集数据与 IRT 模型的拟合程度,运行 ConQuest 软件可得到两类卡方拟合指标: Outfit MNSQ (Outfit Mean Square) 和 Infit MNSQ (Infit Mean Square) . 其中 Infit MNSQ 则指代加权后的残差均方,反映了与题目难度水平相当的学生的作答方式是否与模型一致,若偏差大则说明作答模式不一样. Outfit MNSQ 是残差的均方,Outfit MNSQ 对数据中的极端值较敏感,而对于那些项目难度和被试能力水平相当的数据 Infit MNSQ 则表现地更为敏感. Outfit MNSQ 和 Infit MNSQ 可接受的取值范围和研究目的有很大关系,不同的研究目的可接受的范围略有不同. 通常认为 Outfit MNSQ 和 Infit MNSQ 取值范围在 0.75 ~ 1.3 为拟合度良好[8]. 若很多能力强的被试答错题目,而能力弱的被试答对题目,一般残差会比较大,说明该题目不符合 IRT 模型的预期,可能是因为题目表述有问题,导致能力强的个体答错,也可能是该题目与其他题目所测试的潜在心理构造不同,需综合各项指标进行分析.

3.3　怀特图

怀特图又被称为项目—被试图,是根据 IRT 模型将数据进行分析处理后将项目难度与被试能力映射在一个共同量尺上,从而使测量更加直观具体. 通常以一根竖线表示测量值的大小,从下往上依次增大,在竖线的两边分别为项目难度与被试能力. 怀特图以 Logit 为基本单位,将被试能力与项目难度放在同一量尺上的左右两边,把它们视为一个问题的两个方面,可进行直观比较. 通过怀特图,可以判定一个确定水平的被试正确作答某一项目的概率、不同题目之间的难易关系以及试卷的整体难度与被试群体的水平是否适宜等多种信息.

最左边数字表示的是量尺的刻度 logit 单位,在竖线左边的是被试的能力分布,它表明考生所处量尺的位置,竖线右边则是试题的难度分布. 自下往上,试题难度和学生能力值依次升高,即在怀特图最下面的学生的能力水平较低,试题的难度最小较为简单. 学生之间的差距表示被试能力水平之间的差距,而试题之间的距离表示的是不同试题的难度水平之间的差距. 需要注意的是,在量尺的"0"处表示个体正确回答题目的概率为 50%,能力在 0 以上的学生对 0 以下的项目正确回答的概率大于 50%. 以下图为例,怀特图的检验主要关注三个问题: 被试能力范围分布; 试题难度范围分布; 被试能力分布于试题难度分布的匹配程度以及分布形态. 由图可知,试题难度分布约为 4 个 logit 单位,被试能力分布为正

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

态分布且占据约 4 个 logit 单位,试卷整体难度分布跨度与被试能力值分布大致相当,这说明试题的内容覆盖了所有能力水平的考生. 整体来看,试题与被试的能力分布对应较好.

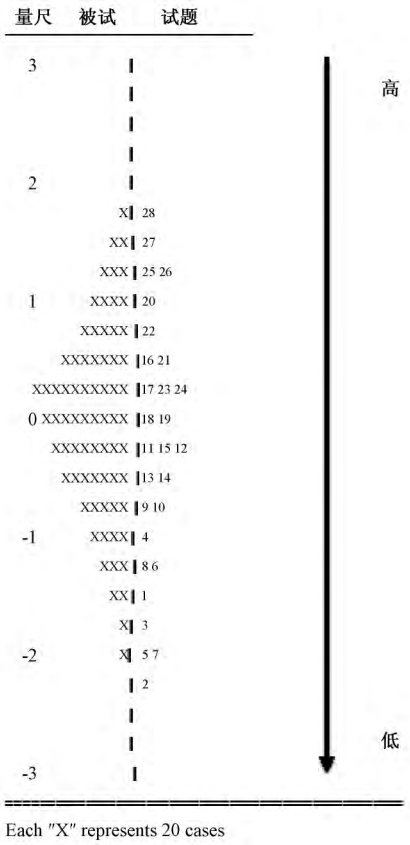


图 1 怀特图示例

3.4 试题信息量分析

试题信息量分析分析的基本思想是: 设计良好的试题应能为测试提供较多的信息,降低对被试能力水平估计方面的误差. 项目反应理论认为,用与被试能力水平相当的试题进行测试时,试题才能提供最精准的测量结果. 信息量越大的试题对被试的估计越为精确、误差越小,表明该试题在评价该被试的特质水平上越有价值,因此工具检验时应找出信息量较大的试题. 试题的信息量与试题的难度和区分度紧密相关,因此透过试题信息量还可帮助获取试题难度及区分度方面的信息,便于对试题进行改进.

在 IRT 测量中,使用试题信息函数分析试题信息量. 试题信息函数显示出每一道题目对不同能力的被试提供不同的测量精准度. 当题目的难度越接近被试的能力值时,就能提供较高的测量精准度,反之则获得的测量精准度较差. 相同的一份测验对于不同水平的被试,其信息量(测量的标准误差)不同. 试题信息函数图通常有一个峰值,表示试题的信息量达到最大值时,则说明测验对此能力水平的测量误差最小. 即当被试能力处于峰值时,试题可提供最大信息量,该试题适合测试能力值处于峰值附近的被试.

图 2 的横轴坐标表示被试的能力估计值,纵轴坐标表示的试题的信息量. 该图显示了试题在不同能力点上的测量精准度,信息量越高表示试题对该能力点的测量精准度越高. 从图中可以看出,该试题对估计能力在 -0.1 左右的被试时,测量误差最小.

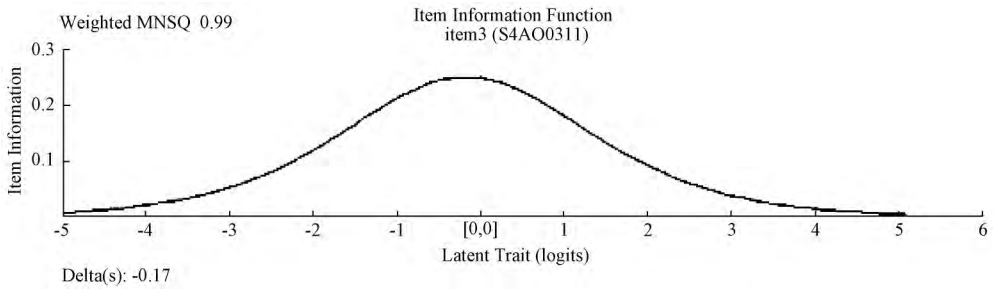


图 2 试题信息函数示例

需要特别注意的是,研究人员若依据试题信息函数验证试题时,首先需保证项目特征曲线能够与试题相拟合,若拟合度差,则会产生误导作用.

4 基于 CTT 和 IRT 的试题质量分析举例

依据本文中试题质量分析的各个指标,以某次小学科学学业质量测试试卷为例,内容覆盖生命科学、物质科学和地球科学 3 个学科领域,对其试题质量进行分析. 参加本次测试的学生人数为 683 人,共计 25 个选择题,2 道问答题,满分为 100 分.

4.1 基于 CTT 的试题质量分析

4.1.1 试卷的总体描述性和整卷信效度分析

本次测试学生平均分为 62.51,标准差 12.65,最高分 92 分,最低分 18 分,偏度 -0.454,考生成绩集中分布在 50 分 ~80 分(54.1%),说明其数据分

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

布形态与正态分布相比略为负偏态分布，即数据左测长尾更长，总体均值在众数的左边，表明被试总分较高，试题难度略易. 峰度为 0.091，说明被试总分分布曲线与正态分布相比更为陡峭，为尖顶峰，表明被试数据分布较为密集地分布在本卷平均分周围. 该卷信度分析采用同质性信度中的科隆巴赫系数(Cronbach's Alpha)，信度为 0.69，说明信度良好. 效度检验采用内容效度分析，通过相关领域的专家根据学科内容和测试蓝图进行逻辑分析后，认为试卷的内容效度良好.

4.1.2　试题难度、区分度分析

本套试卷共计 35 个小题，项目的难度是所有考生在该项目得分的平均数与该项目满分的比值; 区分度是考生在此项目得分与其测验总分间的相关，难度与区分度的散点图见图 3.

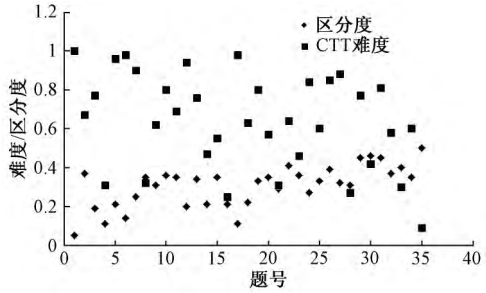


图 3　经典测验理论计算得到的项目区分度和难度散点图

由图可知: 大多数试题的区分度分布在 0.2 ~ 0.4 之间，说明区分度良好，试题的难度分布在 0.2 ~ 0.8 之间，多数试题的难度都分布于 0.5 ~ 0.8 之间，说明试题难度分布较广，整体试卷难度适中.

4.2　基于 IRT 的试题质量分析

4.2.1　试题难度、拟合度分析

用分步评分模型同时对试题中的所有项目进行分析，其结果主要包括试题的 IRT 难度以及 Rasch 模型的 INFIT 拟合度，IRT 难度与 INFIT 拟合度的散点图见图 4.

由图可知，试题难度介于 -5.66 ~ 1.582 之间，Infit 拟合度介于 0.91 到 1.11 之间，所有试题均非常接近于 1，说明所有试题与假设 Rasch 模型有较好拟合.

4.2.2　怀特图分析

在项目反应理论中，被试的能力水平与项目的难度是处于一个共同的量尺之上，因此它们之间可以直接进行比较. 由图 5 可知，考生的能力主要分布在(-2,2) 之间，其中大多数考生的能力水平处于

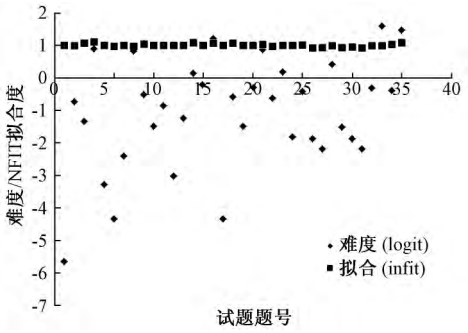


图 4　项目反应理论计算得到的项目难度和 INFIT 拟合度散点图

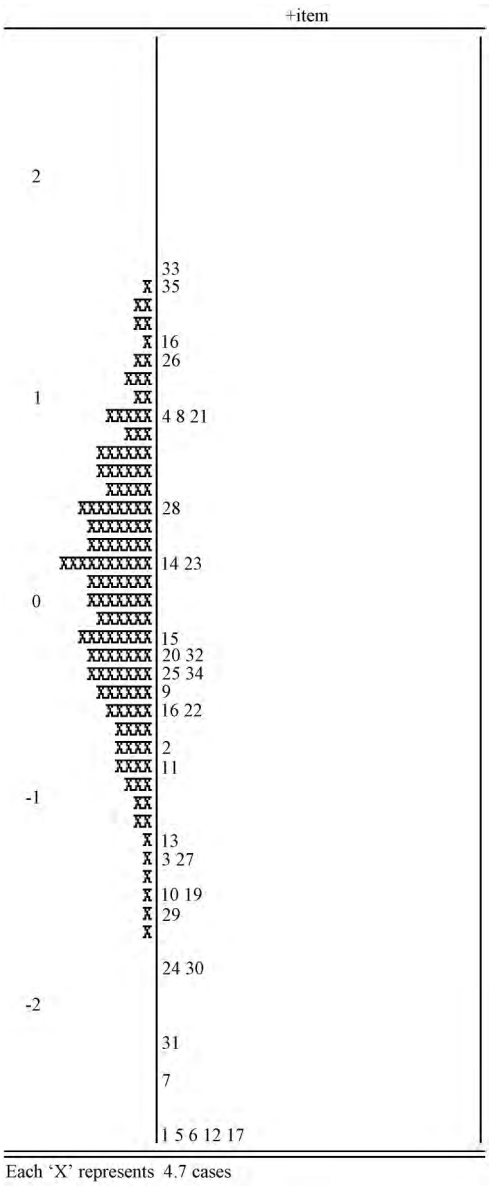


图 5　项目反应理论怀特图分析

(-1,1) 之间. 试题的分布范围较之学生能力的分布范围更广，能较好地涵盖所有学生的能力水平，即学生能力值与试题难度分布匹配性较好.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net

4.2.3 试题信息量分析

项目反应理论提出了测验信息量和项目信息量的概念. 测验的信息量越大,则测验测量能力水平的误差就越小,它与测量标准误差成反比. 相同的一份测验对于不同水平的考生,其信息量(测量的标准误差)不同. 当测验在某个能力水平的信息量达到最大,则说明测验对此能力水平的测量误差最小. 图 6 给出该次测验的信息函数,由图可知,该试卷对估计能力在 1.0 左右的被试时,测量误差最小.

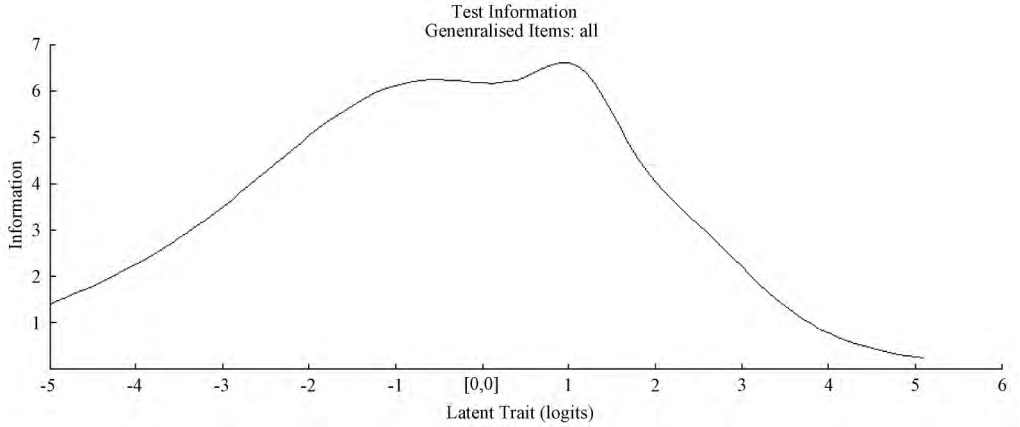


图 6 测验信息量图

5 小 结

为了获得对学生知识和能力等的科学客观诊断结果,提高测试的信效度,需要通过项目分析对试题进行筛选和甄别. 而参考 CTT 和 IRT 的测试结果,可有效帮助研究者提高工具质量,研究者可从以下三方面分析工具质量,适当修改工具以提高工具质量.

1) 从试题个体和工具总体两方面综合考量工具质量. 工具质量的分析不但要考虑各个单题的质量,也要考察工具整体的信效度,从而找出影响工具质量提高的关键要素和单题. 其中衡量单题质量的主要技术指标包括: 试题难度、区分度、选项分析. 在工具整体层面的分析主要涉及信度和效度分析. 综合两方面的检验结果,可找出有待修改的试题,从而提高工具质量.

2) 综合 CTT 和 IRT 理论的优势,综合检验试题的质量将进一步提高工具的信效度. 当前通常依据经典测验理论(CTT)和项目反应理论(IRT)综合分析工具测试结果以提高工具质量,其中经典测试理论的检验参数主要包括: 难度、区分度、选项分析等,而项目反应理论则重点关注试题难度、项目拟合度、怀特图、标准误、试题信息量等参数.

3) 统计指标是试题修改的辅助工具,研究者除了综合参考各种统计指标外,试题的修改及删除与否还需要参考试题设计的理论框架和测试蓝图等,结合测试目的才能最终确定试题的修改方向. 工具质量评估的过程是一个不断寻找证据支持论证工具信效度和客观性的过程,除了侧重量化分析的测量学指标的运用,还需特别注意参照工具开发的测试目的以及理论框架,这些均能够为工具质量评估提供重要的证据支持,因此要充分重视并综合运用这些信息,促进高质量工具的开发.

参 考 文 献

[1] 辛涛. 项目反应理论研究的新进展[J]. 中国考试,2005(07):19.

[2] 雷新勇. 考试数据的统计分析和解释[M]. 上海: 华东师范大学出版社,2007:50.

[3] 雷新勇. 大规模教育考试: 命题与评价[M]. 上海: 华东师范大学出版社,2006:232.

[4] Jack R, Fraenkel, Norman E. Wallen, Helen H. Hyun. How to design and evaluate research in education[M]. New York: McGraw-Hill,2012:154 – 159.

[5] 余民宁. 心理与教育统计学[M]. 台北: 三民书局,2005: 494.

[6] 约克奇 R D. SPSS 其实很简单[M]. 北京: 中国人民大学出版社,2010: 60.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved. http://www.cnki.net

[7] 袁克定. SPSS for Windows 数据统计分析工具应用教程 [M]. 北京: 北京师范大学出版社, 2001: 277.
[8] Bond T G, Fox C M. Applying the Rasch model: fundamental measurement in the human sciences, 2nd ed [M]. New York, East Sussex: Routledge, 2007: 238 - 241.

Integrating CTT and IRT to Improve the Quality of Monitoring Tools ——Taking Primary Science Examinations as an Example

Zhou Gaixiao[1,2]　Liu Enshan[2]

(1. College of Life and Environmental Science, Wenzhou University, Wenzhou Zhejiang　325035;
2. College of Life Science, Beijing Normal University, Beijing　100875)

Abstract

The test items are the commonly tools in the science academic quality assessment, and they are also an important tool for carrying out educational empirical research. The quality of the tool directly determines the reliability and validity of academic assessment and educational research. Based on the classical test theory (CTT) and the item response theory (IRT), the paper analyzes the conventional process and methods of the quality analysis of the test items, the connotation of the relevant parameters of the tool quality assessment is explained to help the test developers and educational researchers understand the quality analysis methods and techniques of test items, so as to effectively improve the quality of academic evaluation and educational research tools. It will help to improve the reliability and validity of the academic quality evaluation and the educational research.

Key words: classical test theory, item response theory, test paper, quality evaluation.

(C)1994-2020 China Academic Journal Electronic Publishing House. All rights reserved.　http://www.cnki.net