两种教育测量理论在试卷质量控制和评价中的应用及其展望

钟 轶[1],季晓辉[2]
(1.南京医科大学公共卫生学院,2.党委组织部,江苏 南京 210029)

摘 要:研究归纳了经典测量理论与项目反应理论的异同,通过文献检索的方法阐明项目反应理论在国外教育机构已得到普遍运用,而国内医学院校的试卷质量控制和题库建设还停留在经典测量理论阶段,需要指导理论和方法的革新。

关键词:教育测量;经典测量理论;项目反应理论
中图分类号: G642.47 **文献标志码**: A **文章编号**: 1671-0479(2013)01-066-006
doi:10.7655/NYDXBSS20130118

教育测量是评价教育结果的重要路径之一,而考试又是教育测量的重要方法之一。教育测量不仅仅是对考试结果的评价分析,也包括了对试题试卷质量的评价分析。随着各种考试深入到社会生活的各个领域,试卷质量成为现代考试组织管理者需要关心的核心问题。既要保证考试对所有应试者公平、公正,又要能较好地考核应试者的能力,以利选拔人才。这就需要使考试的各个项目参数保持在一个相对稳定的层面上。教育测量学理论为如何获得项目参数,如何运用参数进行长效管理提供了理论依据。目前人们主要利用两种测量理论——经典测验理论(classical test theory, 简称 CTT) 和项目反应理论(item response theory,简称 IRT)来进行试卷质量控制和评价。考试管理者需要对这两种理论进行深入的了解并熟练掌握,才能对每一次考试形成正确的评价。

一、教育测量的定义及其历史发展

教育测量是依据一定的教育学理论,使用测验对人的教育成绩进行定量描述的过程[1]。世界上最早的教育测量出现于中国西周奴隶制时期(公元前 1100 年~公元前 771 年)。隋炀帝大业二年 (公元 606 年)出现的科举制在中国延续了 1 300 年。其间创造的分科考试、“弥封”、复评等方法,在我国早期教育测量方面产生了较大的影响。

现代教育测量的理论与技术产生于工业革命后的一些西方国家, 从 20 世纪初叶开始获得迅速发展。20 世纪 50 年代以后经典测验理论趋于成熟并稳步发展,得到广泛应用,在相当长的一段时期里成为指导教育测量的主流方法和核心理论。20 世纪 60 年代以后出现了 IRT 和概化理论(generalizability theory,简称 GT),并逐渐打破了 CTT 的一统天下,引起了该领域的深刻变革并影响至今。在当今西方国家,IRT 指导下的教育测量学理论与方法得到迅速发展与广泛应用, 在许多方面已呈现逐步取代 CTT 的态势。

二、两种教育测量理论的概念、特点及其应用

(一)CTT

1. CTT 简介及优缺点分析

CTT 是基于 E. L. Thorndike 的 “凡客观存在的事物都有其数量”和 W. A. McCall 的“凡有数量的东西都可以测量”这一可测性假设提出的。其基本思想是把测验的得分 (通常称为测验的观察分) 看作真分数(反映被试者)和误差分数的线性组合,其数学模型表示为:$X = T + E$。其中 X 是观测分数,T 是真分数,E 是随机误差。1950 年,Harold Gulliksen 根据这一模型,引申出 3 个相关联的假设公理:① 若一个人的某种心理特质可以用平行的测验反复测量足够多次,则其观察分数的平均值会接近于真分数。

收稿日期:2012-09-18
作者简介:钟 轶(1981-),男,江苏常州人,南京医科大学公共卫生学院 MPH 学员,研究方向为高等医学教育与考试管理。

② 真分数和误差分数之间的相关为零。③ 各平行测验上的误差分数之间相关为零。

对于利用 CTT 进行分析的测试而言,一旦测试结束,即可利用测试结果进行分析。其中重要的参数指标为信度 (reliability)、效度 (validity)、难度(difficulty)和区分度(discrimination)。信度用以衡量测验结果是否反映了被测者的稳定的、一贯性的真实特征。在 CTT 中被定义为一组测量分数的真分数的方差(变异数)在总方差(总变异数)中所占的比率。实际运用中是计算反映试卷内在一致性的 Cronbach' α 值,如大于 0.7,可认为信度较高。效度(validity)即有效性,是指测量工具或手段能够准确测出所需测量的事物的程度,测量结果与要考察的内容越吻合,则效度越高;反之,则效度越低。CTT 对效度问题提出了诸多解决方案, 美国心理学会在 1974 年将测量的效度分为三大类, 即内容效度(content validity)、结构效度(construct validity)和效标关联效度(criterion-related validity),以检测知识为主的考试较容易获得较高的内容效度。

难度和区分度属于项目分析的范畴,目的是筛选、甄别项目以提高测验的信度和效度。难度的主要指标是难度系数(也称通过率,标记为 P),即在该题上答对的人数与全体被试的比率(或平均得分与该题满分的比率)。难度等级可按 P 值划分:P <0.6 为难,$0.6 \leqslant P < 0.8$ 为中等,$P \geqslant 0.8$ 为简单, 一份试卷的合理难度应介于 0.7~0.8。仅难度还不足以说明题目质量的优劣,CTT 还提出区分度概念。区分度(标记为 D)是指一道题能多大程度上把不同水平的人区分开来,D 值越高,越能把不同水平的受测者区分开来,该道题目被采用的价值也就越大。一般认为 D 值 0.40 以上为优良,0.19 以下为差应淘汰。

CTT 的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。它是测验中最一般、最基本的理论,曾在国内外的心理与教育测量中被广泛应用。但因为 CTT 存在着难以克服的技术问题,在现代测量理论诞生后,其在教育测量中的地位有所下降,在应用方面被 IRT 超过。CTT 的缺点主要是:①真分数的方差和误差分数的方差无法获得,而严格意义的平行测验又无法实现,使信度只能估计,不能直接计算,导致估计精度低。② 各种参数估计严重依赖样本,难以避免抽样误差。③ 用一个误差指标 SE 来描述所有测试的测量精度, 过于笼统和单一。④ 试题难度和被试者水平如果不在同一个参照系上,依靠现有的参数指标,找不到验证某试题是否恰好匹配某被试者(或群)的计量方法。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。

2. CTT 在国内外教育测量中的应用情况

目前国外已很少单独采用 CTT 进行测量数据分析和评价, 多为和 IRT 联合使用。Jessica Sharkness 和 Linda DeAngelo 利用 CTT 和 IRT 从社会参与和学术参与两个维度对加州大学(UCLA)2008 份"你的大学第一年"调查数据进行了分析,指出相比较 CTT,IRT 不但在测量精度方面能提供更为丰富的信息, 而且在质量改善上能提供更明确的路线图[2]。

与国外相比,IRT 在国内的影响和应用还处于刚刚起步阶段。由于对 IRT 了解不够,目前国内医学院校更倾向于使用 CTT 对医学生在校期间的医学课程考试情况和试卷质量进行分析和归纳整理。自 2000 年以来主要有:首都医科大学燕京医学院分别对三年制卫生信息管理专业手术分类学期末考试试卷[3]和七年制临床医学专业组织学与胚胎学期末试卷进行质量分析[4];重庆医科大学对 2006 级临床医学专业本科生外科学期末考试试卷进行质量分析[5];南京医科大学对 2003~2005 级五年制临床医学专业学生的病理学、药理学、妇产科学、内科学的试卷进行了质量分析 [6]; 福建医科大学对 2005 级预防医学专业学生流行病学考试成绩进行综合分析[7];北京大学医学部对 2001 级八年制临床医学专业学生医学细胞生物学期末考试试卷进行了试卷的效度、信度、难度、区分度和试题的难度、区分度分析[8];大连医科大学对2001 年度机能实验考试进行了成绩分析[9];南京中医药大学对 225 份生理学期末考试试卷进行分析[10]等。

通过这些分析, 各医学院校对本校的试卷质量有了直观的认识。认为试题难度适中,区分度较好的有 3 所学校;认为试题难度偏低或偏高,区分度较差的有 2 所学校;另有 2 所学校未给出整体评价。各院校均认为有必要对试卷质量进行持续监控, 并不断提高。但由于被试对象和各个题目参数的不同,这些分析无法进行校与校之间的比较, 使分析的普遍性受到局限。而且就试卷质量分析而言,即使在同一学校范围内,若按照 CTT,只要被试者正确回答的题目数量相等,那么被试者的能力也相等。但实际上每道题目所包括的知识点都是不相同的, 被试者正确回答题目数量相等不能等同于被试能力相等, 应用 CTT 进行分析掩盖了这种不同。同时各校也没有再提供后续的试卷质量分析, 无法得到试卷质量是否得到改善的数据,使分析的实用性下降。

CTT在国内医学院校的另一个主要应用是和计算机技术相结合来组建题库和考试系统。目前组建题库有两个方向:单机版和网络版,如内蒙古医学院组建了基于网络平台的医学微生物试题库(网络版)[11];重庆医科大学附属第二医院组建了外科学题库并自主研发了自动组卷系统(单机版)[12];第三军医大学构建临床医学题库管理系统,使临床教学完全实现了教考分离的目标(网络版)[13];天津医学考试中心建立了基于VB6的计算机辅助医学考试系统(网络版)[14];南方医科大学自主研发了基于微软SQL2005数据库平台的智能网络题库与考试系统(网络版)[15]等。以上各种题库的构建为组织标准化考试、实现教考分离、提高试卷质量进行了有益的探索。特别是基于网络技术,使项目参数易于获得和分析,极大地提高了考试后试卷分析的效率和准确率。可以认为,基于网络化的题库和考试系统是日后题库建设的趋势。

同时应注意到,现代技术的引入并没有解决CTT自身的缺陷。以上题库是构建在不同的项目参数上的,相互间很难比较。又因为各参数极度依赖被试样本,被试者(或群)一旦不同,测量结果就会发生变化。应用CTT所得项目参数被局限在一校一院的范围内,无法外推。

综上所述,CTT尽管已经取得了公认的成就,但其在未来的应用会受到限制,需要有新的理论和方法来规避CTT的缺陷,为教育测量提供更为准确的指导理论。

(二)IRT

1. IRT简介及特点分析

IRT是建立在潜在特质理论基础上的。潜在特质是指被试者不能被直接观察到的某种稳定的、支配其对相应的测验项目做出反应,并对反应表现出一致性的内在特征(记为θ)[16]。被试者的某个潜在特质与测量该特质的项目反应之间存在着如下关系:随着潜在特质θ的提高,正确反映该项目的概率$P(\theta)$也提高。IRT是研究θ与$P(\theta)$之间的函数关系,并用一定的数学模型来反映两者关系的一种测量理论[17]。IRT有各种各样的模型,目前应用较多的是二级评分模型中的单参数Logistic模型(简称1PL模型或Rasch模型)、双参数Logistic模型(简称2PL模型)和三参数Logistic模型(简称3PL模型),3PL模型公式为:

$$P_i(\theta)=c+(1+c)\frac{1}{1+e^{-Da_i(\theta-b_i)}},i=1,2,\cdots,n$$

公式中a、b、c分别对应了项目区分度、项目难度、猜测系数这3个参数,D为常数1.7。如不考虑猜测因素,则c=0,模型变为2PL模型;如再假设所有项目区分度一致仅难度不同,则a=0,c=0,模型变为1PL模型。

据此公式,可形成项目特征曲线(item characteristic curve,简称ICC),从而直观地展现项目各参数情况,见图1。ICC横轴为被试者的能力值,纵轴为被试答对该项目的概率,随着被试者能力的提高,被试答对项目的概率也逐步提高。猜测系数c(取值范围0~1)为该曲线的下限,表明即使被试能力很低,也有猜对答案的可能性。难度系数b(取值范围-3.00~3.00)表示在曲线最陡的那一点所对应的θ值,b值变大,曲线向右发生平移,但形状不变,揭示项目难度变大,需要被试者有更高的能力才可能达到原来的应答概率。反之,b值变小,曲线向左发生平移,揭示项目难度下降。区分度a(取值范围0~3.00)为曲线拐点处的斜率,即斜率的最大值。a值变化会引起整个曲线的形状发生变化。a值变大,则曲线愈加陡峭,表明项目区分度较好;a值变小,曲线变平坦,表明项目区分度不佳。

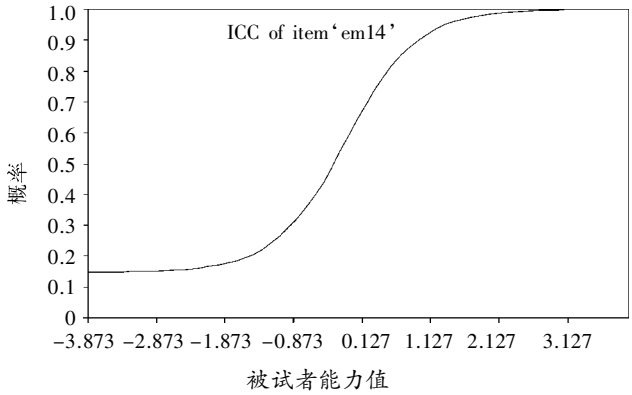

图1 项目特征曲线

相较于CTT建立在弱假设基础上,IRT是建立在强假设基础上的,其基本假设为:① 测验的潜在空间(latent space)的单维性假设,即组成某个测验的所有项目都是测量同一潜在特质。② 测验项目的局部独立性假设,即对某个被试能力而言,项目间不存在其他任何关系。③ 测验的ICC模型假设,即在IRT中,被试者对一个项目正确反应的概率$P(\theta)$是由所假设的IRT模型、项目参数以及被试相应的能力或特质的水平θ所决定,而与被试群总体的能力分布无关。只要找到适合数据的模型,就可以对项目进行比较精确的分析。

IRT的优点主要是:① 采用局部独立性假设与样本独立项目校准的方法,解决了CTT对样本依赖

性大的问题。② IRT 项目参数具有不变性,各被试者(或)群所得的项目参数具有可比性。③ 通过 ICC 可以预测被试者在一个新测验项目上的正确反应概率 $P(\theta)$。④ 在题库建设中,在难度、区分度和猜测系数的基础上增加了题目信息函数,从而提高了题库参数的完备性和题库管理的可控性。

除以上优点外,IRT 也因为自身一些缺点而使其应用受到限制。比较突出的一个缺点就是其数学模型复杂,计算工作量大,单纯依靠手工很难完成。其次 IRT 估计参数时还是必须通过测试获得,不同的被试者,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,就很难检测出数据与模型间存有的偏差。

2. IRT 在国内外的应用情况

IRT 诞生后, 学界对其产生了极大的热情和兴趣,其理论和应用不断得到充实和扩展。

在欧美发达国家,在教育领域应用 IRT 进行统计分析、质量监控、题库建设以及考试设计已经常态化。美国国家教育进展评估(NAEP)的实施与结果报告即通过把一系列复杂的抽样设计与 IRT 的模型方法相结合,最终生成反映学生总体熟练程度的各种统计量[18];荷兰通过基于 IRT 的监测与评估系统对全国初等教育教学质量进行监控并形成报告,荷兰学校利用这些报告改进和提高教学水平[19];美国医师执照考试(USLME)也应用 IRT 指导各项工作,美国国家医学考试委员会 (NBME) 选用单参数 Logistic 模型进行数据统计与分析,以保证这项几乎天天都进行的考试在各项目参数上保持一致[20]。

在国内,理论方面,《考试研究》期刊系统地将美国亚利桑那州立大学教授 Joseph. M. Ryan 博士的《基于经典测量理论和项目反应理论的等值与连接》介绍到国内 [21-23], 此为美国教育测量学会(NCME) 2010 年年会培训课程内容。应用方面,首都医科大学燕京医学院提出将 IRT 用于高校标准化题库的建设[24];上海教育考试院专门研究了 IRT 在评价考试命题质量中的应用问题,认为在大规模教育考试中应用 IRT 对命题质量进行评价,有着重要意义[25];东南大学对 1987 年上海市不同高校入学考试数学试题中前 12 道填空题进行了基于 IRT 的试卷质量分析[26];北京语言大学以实测 5 890 份初、中等汉语水平考试(HSK)考生数据为母本,专门对 IRT 中的三种估计考生能力值的方法进行比较,结果表明能力值估计结果与考生潜在能力分布有关系。当潜在能力分布趋向正态分布时,能力值的估计误差较小。此外,不同软件的参数估计方法的能力值估计结果均有差异[27];广州市教育局教研室、华南师范大学心理应用研究中心合作的“基础教育教学质量监测系统”项目组,探讨了 IRT 下试题数量和被试量对参数估计的影响,认为在两级试题模拟测验情境下,随着被试量和题量逐渐增大, 项目参数估计值模拟返真指标均方误差逐渐减小[28]。广州大学、华南师范大学的学者以广东省佛山市中考数学实测数据为例,对 IRT 测验等值模型的选择过程进行了说明[29];清华大学教育研究院和江西师范大学自主研发了混合模型参数估计程序 Mix-Tu,该程序具有较高的返真性,与国际知名的 IRT 分析软件 Parscale 相当[30]。

此外,基于 IRT 的计算机化自适应测验(CAT)系统的研发受到越来越多的关注。教育部考试中心[31]、贵州师范大学[32]、云南师范大学[33]、上海交通大学[34]、天津师范大学[35]等机构从多个角度提出了构建CAT系统方法及评价标准。这为 IRT 日后的推广应用打下了基础。

当然也应看到,IRT 在国内应用的广度和深度与欧美国家相比有着明显的差距。理论和应用创新较少;多数教师和教育管理者对 IRT 尚不了解。使用 IRT 指导出卷和题库建设的国内高校还较少;基于 IRT 的试卷分析报告在时间上缺少延续性,孤立分析报告的实用度受到限制; 目前国内也无大型考试采用 CAT 系统, 已经建成的系统还存在着重开发、轻推广利用的问题。国内高校要利用好 IRT,还需要相当长的时间。

三、总结与展望

通过应用教育测量理论, 高校教育管理者可以获取每次考试的相关项目参数, 再对这些项目参数进行归纳整理, 以实现考试质量控制的常态化和促进教学质量的提高这一目标。

由于各具特色,CTT 和 IRT 将在较长的时间内共存下去。但 IRT 的应用较多的趋势不会改变。理论界对 IRT 的补充和完善也不会停止,多维项目反应理论将有可能是下一步研究内容之一。

国内高校和考试机构将进一步完成数据积累和理论学习,以实现基于 IRT 的高质量题库建设和大规模应用。在日本,各医学院校间共用考试系统已经完成, 由各医学院校联合对学生进行高质量的统一考试[36]。这对国内如何保证医学生培养质量是一个很好的启发。

南京医科大学于本世纪初开始在四年级临床医学专业培养中引入教考分离制度,对内、外、妇、儿四

门主要课程由教务科通过题库统一命题和组织考试，在保证考试公平性和促进各见习医院教学质量改善上做了有益的尝试。但因为题库是建立在 CTT 基础上的，CTT 的固有缺陷也被完整地继承到了统考制度中。就考前题库抽题和组卷来说，试题的项目参数在考试后实际上会发生变化，应该给予修正，但题库是不能修改的封闭的系统，固定项目参数的试题实际上降低了考试的信度和效度。就考试后的数据分析来看，尽管每次考试的项目参数可以获得，但由于被试对象不同，考试之间的项目参数缺乏可比性，难以对每一届学生的培养质量进行比较。所以，尽快将 IRT 引入学校的试题库建设和考试组织，应该被提上议事日程。

应该清楚地认识到，南京医科大学对于现代教育测量学的认识和运用目前还处于较初级的阶段，学校缺少专门的考试管理部门——考试中心；从管理部门到教研室对于考试普遍存在着重组织轻分析的问题；教研室对教考分离和题库建设的积极性不高；学生以应付心态应对考试的占大多数。

世界医学教育联合会推荐的本科医学教育的国际标准，将对学生的考核高标准定义为：考核方法的信度和效度应当有评定并记录在案，不断开发新的考核方法。因此，南京医科大学医学生培养要达到国际标准，就必须真正掌握现代教育测量理论，通过不断引入先进的方法，形成稳定、科学、高效的考试管理制度并有效运用，从而实现与国际真正的接轨。

参考文献

[1] 戴海崎，张 锋，陈雪枫. 心理教育测量[M]. 广州：暨南大学出版社，1999：10

[2] Sharkness J，DeAngelo L. Measuring student involvement：A comparison of classical test theory and item response theory in the construction of scales from student surveys [J]. Research In Higher Education，2011，52(5)：480-507

[3] 蔡纳新，鲁 杨，聂立华，等. 卫生信息管理专业手术分类学期末考试试卷分析与评价[J]. 现代医药卫生，2012(4)：621-622

[4] 翁 静，郭晓霞，路 欣，等.《组织学与胚胎学》试卷分析及其参考意义 [J]. 首都医科大学学报：社会科学版，2010，14：222-226

[5] 李增志，汪 林，黄小明，等. 应用自动组卷系统组卷对外科学课程期末考试实施教考分离的试卷分析 [J]. 现代医药卫生，2012 28(1)：136-137

[6] 韩春红. 对南京医科大学临床医学生部分课程考核试卷、试题质量的分析[J]. 南京医科大学学报：社会科学版，2010，10(1)：63-69

[7] 蔡 琳，许能锋，何保昌，等.《流行病学》试题库的应用与分析[J]. 福建医科大学学报：社会科学版，2010，11(3)：30-32

[8] 宋 青，肖军军. 医学细胞生物学试卷考核质量评价[C]//中国细胞生物学学会 2005 年学术大会、青年学术研讨会论文摘要集，2005：46

[9] 邢 嵘，贾玉杰，康晓楠，等. 机能实验期末考试试卷分析及评价[J]. 医学教育，2002(3)：34-35

[10] 陈 明. 对 225 名学生生理学期末考试的分析与思考[J]. 西北医学教育，2000(2)：86-88

[11] 陶格斯，张明昱，卢 莎，等. 利用教学资源网平台构建医学微生物学试题库探讨[J]. 基础医学教育，2012(3)：233-235

[12] 李增志，汪 林，黄小明，等. 应用自动组卷系统组卷对外科学课程期末考试实施教考分离的试卷分析[J]. 现代医药卫生，2012(1)：136-137

[13] 李 由，刘光琼，向国春，等. 教学医院题库管理系统建设及应用 [J]. 重庆师范大学学报：自然科学版，2011，28(6)：49-51

[14] 孙建华. 基于 VB6 的计算机辅助医学考试系统的研究[J]. 医学信息，2011(7)：4063-4064

[15] 耿景海，席卫文，张春辉，等. 医学网络题库与考试系统利弊分析——基于网络考试的事实经验[J]. 西北医学教育，2012(1)：159-161

[16] 顾海根. 应用心理测量学[M]. 北京：北京大学出版社，2010：266

[17] 吉尔伯特·萨克斯. 教育和心理的测量与评价原理[M]. 南京：江苏教育出版社，2002：322

[18] Thissen D，王丽华. 国家教育进展评估的效度研究[J]. 考试研究，2012(2)：66-76

[19] Moelands HA，王丽华. 荷兰初等教育监测与评估系统[J]. 考试研究，2011(6)：3-12

[20] 张 鸣，万学红. 心理测量理论与技术在美国医师执照考试中的应用[J]. 复旦教育论坛，2003(2)：88-90

[21] Ryan JM，杜承达. 基于经典测量理论和项目反应理论的等值与连接——主要概念和基本术语[J]. 考试研究，2011(1)：80-94

[22] Ryan JM，杜承达，谢小庆. 基于经典测量理论和项目反应理论的等值与连接——等值设计和经典测量理论等值程序[J]. 考试研究，2011(2)：83-95

[23] Ryan JM，杜承达，谢小庆. 基于经典测量理论和项目反应理论的等值与连接——项目反映理论等值程序[J]. 考试研究，2011(3)：80-94

[24] 李 丹，刘春华，董建鑫. 基于项目反映理论的高校标准

化试题库建设的探讨[J]. 数理医药学杂志,2011(6):754-755
[25] 王晓华,剑 冰. 项目反应理论在教育考试命题质量评价中的应用[J]. 教育科学,2010(3):20-26
[26] 杨 亮. 基于项目反映理论的试卷质量分析[J]. 长春大学学报,2011,21(2):64-67
[27] 马洪超. 基于 IRT 不同参数估计方法的考生能力估计结果的比较[J]. 考试研究,2012(1):61-66
[28] "基础教育教学质量监测系统"项目组. IRT 下题量与被试量对参数估计模拟返真性能的影响 [J]. 中国考试,2009(6):3-10
[29] 黎光明,张敏强. IRT 测验等值模型的选择—以广东佛山市中考数学实测数据为例[J]. 中国考试,2012(2):8-13
[30] 涂冬波,蔡 艳,戴海琦,等. 项目反应理论新进展:基于 3PLM 和 GRM 的混合模型[J]. 心理科学,2011(5):1189-1194
[31] 关丹丹. 纸笔考试与计算机自适应考试的等效研究探讨[J]. 中国考试,2011(10):13-16
[32] 杨建原,柏 桧,赵守盈. 计算机自适应测验开发的程序研究[J]. 中国考试,2012(3):3-7
[33] 杨跃诚,钟汝能,孙 瑜,等. 基于 IRT 的计算机化自适应测试系统研究[J]. 云南大学学报: 自然科学版,2011,33(S2):294-298
[34] 姚敏华,王 东. 基于项目反应理论的影响计算机化自适应测试 a 分层法的因素研究[J]. 计算机应用与软件,2011(10):149-154,173
[35] 张 滨. 对当前在线考试系统存在问题的分析与改进建议[J]. 佳木斯教育学院学报,2011(3):310-312
[36] 武玉欣,郝素彬. 日本医学教育中的标准化·核心·课程与共用考试介绍[J]. 日本医学介绍,2005(7):332-333

Application and development of two kinds of education test theory on examination paper quality control and evaluation

Zhong Yi[1], Ji Xiaohui[2]
(1.School of Pubic Health, 2. Organizing Department of CPC Committee, Nanjing Medical University, Nanjing 210029, China)

Abstract: This study summarized the similarities and differences between the classic test theory and the item response theory, illustrated that the illustrate item response theory in foreign education institutions had been widely applied while the examination paper quality control and test construction in domestic medical colleges and universities still relied on classical test theory through the method of literature retrieval, and then advanced the necessity of the innovation of theory and methods.

Key words: education test; classic test theory; item response theory