

义务教育阶段学业质量监测题库建设的理论与实践

田 一 张咏梅

(北京教育科学研究院 北京 100191)

摘 要:义务教育阶段学业质量监测是学生学业质量评价的重要方式,在当前国内外教育质量评价领域中发挥着极为重要的作用。加强学业质量监测的题库建设,则是促进学业质量监测工具的命制与实施从经验型向科学化迈进的必由之路。本文在对义务教育阶段学业质量监测题库的性质和功能进行阐述的基础上,重点探讨了学业质量监测题库建设的理论基础、基本流程和关键环节,并进一步分析指出了当前题库建设中有待深入研究的问题。

关键词:义务教育阶段;学业质量监测;题库建设

中图分类号: B841.2

文献标识码: A

文章编号: 1003-5184(2014)01-0062-06

题库即试题库,是按照一定的教育测量理论,在计算机系统中实现的学科试题的集合,是在数学模型基础上建立起来的教育测量工具(余胜泉,朱凌云,2000)。题库的出现是教育测量技术与计算机技术共同发展的结果,是教育考试管理与实施走向规范化、专业化、科学化的重要途径与方式。通常,高级题库拥有存储、查询、组卷、在线实施等功能,例如,由美国教育考试服务公司开发的 GRE、TOFEL 题库,香港基础能力学生评估项目的中央试题库等(香港考试及评核局,1977)。题库由于在管理方面具有高效、灵活、保密的特点,有利于提高命题的科学性与试卷质量的稳定性,已成为当前国内教育测量领域的研究热点与考试评价领域的实践发展方向。

学业质量是教育质量评价的核心指标,跨区域的大规模学业质量监测是评价学业质量的重要方式。北京市义务教育阶段大规模学业质量监测(以下简称学业质量监测),是依据国家课程标准、不同版本教材使用情况与义务教育教学的实际情况,对于义务教育阶段学业质量进行判断的标准参照测验,已有将近十年的发展历程,在学业质量评价领域积累了较为丰富的理论与实践经验,也积累了大量经过测试、具有实测指标的学科试题集。随着对于测验质量要求的不断提高、实施次数的不断增加,改造提升义务教育阶段学业质量监测试题与试卷质量、提高管理与实施水平已成为当前的工作重点。在此背景下,借鉴国内外相关研究领域经验与成果,建设义务教育阶段学业质量监测题库已成为当前的重要举措。

1 义务教育阶段学业质量监测题库的性质与功能

在教育测量学中,根据性质可以将测验划分为常模参照测验与标准参照测验两大类。就北京市义务教育阶段学业质量监测而言,其依据国家课程标准及测试方案中所明确界定的学习任务范围、相对固定的学业标准来解释学生群体或个体学业表现,本质上属于标准参照测验的范畴。同时,还要求具有一定的诊断性功能。义务教育阶段学业质量监测的性质决定了所建题库的性质,监测题库中的每道试题均应与课程标准中的内容标准、与学业水平相应等级标准有明确对应,且自动生成的试卷应具有标准参照测验的性质,并为学生、教师、学校提供相应的诊断反馈和教与学改进的指导服务。

根据义务教育阶段学业质量监测的目的与性质,所建题库功能主要体现在:在宏观层面,实现题库的自动组卷功能,为命制学业质量监测的纸笔测验提供最优选择方案,为科学有效地评价义务教育教学质量提供科学的测量工具;在微观层面,通过题库不断更新开发升级,实现基于题库的计算机自适应测验功能,为评价学校、教师、学生的日常学业水平状况提供适时在线的评价、诊断与反馈。

2 义务教育阶段学业质量监测题库建设的理论基础

建设义务教育阶段学业质量监测题库的理论基础主要为教育目标分类学理论、教育测量学理论等。

2.1 教育目标分类学理论

教育目标分类学理论是建立学科测验方案的理论依据,决定着题库中试题、试卷呈现的内容与结构。其中,以 Bloom 等人在 1956 年提出的教育目标分类学理论应用最为广泛,他指出,教育目标是明确

阐述希望通过教育过程使学生得以变化的方式,即学生改变其思维、感情和行为的方式,可以分为认知、情感和动作技能三个领域。认知领域包括有关知识的回忆或再认,以及智力能力和技能形成等方面的目标;情感领域包括描述兴趣、态度和价值形成等方面的目标;动作技能领域包括操作、运动技能形成等方面的目标(布卢姆等,1986)。

目前,学业质量监测题库试题命制时所考查的能力层级的确定主要以认知领域的目标分类作为理论依据,相应划分为知识(Knowledge)、理解(Comprehension)、应用(Application)、分析(Analysis)、综合(Synthesis)、评价(Evaluation)等六个层次,这是保证题库试题能够科学有效考查学生达到相应掌握水平的主要依据。

2.2 教育测量学理论

教育测量学理论中的经典测量理论(Classical Test Theory, CTT)和项目反应理论(Item Response Theory, IRT),是进行题库建设的基础测量理论。两种现代测量理论渗透在等值设计、试题分析与校准、标准制定、自动组卷、计算机自适应测验等环节中,两者相辅相成,互为补充。

经典测量理论,是以真分数理论为核心理论假设的测量理论及其方法体系,也称真分数理论。由Gulliksen在1950年提出,其数学模型为: $X = T + E$ 。其中, X 为观察分数,即实际测量到的分数; T 为真分数,即反映某种特质真正水平的分数; E 为随机误差,即观察分数接近真分数的程度(戴海琦,张锋,陈雪枫,2007)。经典测量理论在真分数理论假设的基础上构建了丰富的概念体系,主要包括信度、效度、难度、区分度等基本概念。其中,难度和区分度在学业质量监测中作为统计属性的两项重要指标被纳入题库中。

项目反应理论,是以潜在特质理论为核心理论,假设考生对试题的反应与考生的潜在特质(即观察不到的潜在能力)有关,并通过试题特征曲线描述考生潜在特质与其做出正确反应概率之间的关系。数学模型根据测量数据的属性可以分为两大类:两级计分模型和多级计分模型。前者以单参数对数模型(1-Parameter Logistic Model, 1PLM)、双参数对数模型(2PLM)和三参数对数模型(3PLM)最为常用;后者主要包括等级反应模型(Graded Response Model, GRM)、名义反应模型(Nominal Response Model, NRP)、评定量表模型(Rating Scale Model, RSM)和分部计分模

型(Partial Credit Model, PCM)(余民宁,2009)。学业质量监测题库中的客观题部分采用三参数对数模型,主观题部分采用等级反应模型进行分析。

三参数对数模型(3PLM),由Lord和Novick在1968年提出。其数学模型为: $P_i(\theta) = c_i + (1 - c_i) \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$ 。其中, $P_i(\theta)$ 表示能力值为 θ 的考生答对试题 i 或在试题 i 上做出正确反应的概率; a_i 、 b_i 、 c_i 分别表示试题 i 的区分度、难度、猜测度参数; D 是一个常数,其值为 1.702; e 代表以 2.718 为底的指数。等级反应模型(GRM),由Samejima在1969年提出,是从Lord的双参数对数模型(2PLM)延伸而来,其数学模型为: $P_{ix}(\theta) = P_{ix}^*(\theta) - P_{i(x+1)}^*(\theta)$, 其中, $P_{ix}^*(\theta)$ 的概念与 2PLM 的模型 $P_i(\theta) = \frac{e^{Da_i(\theta - b_i)}}{1 + e^{Da_i(\theta - b_i)}}$ (字母代表的含义与 3PLM 相同)原理一致。具体含义是指在具有能力值为 θ 的考生在试题 i 中得 x 分的概率(Win J. & Ronald K., 1997)。

3 义务教育阶段学业质量监测题库建设的基本流程与关键环节

3.1 基本流程

就北京市义务教育阶段学业质量监测题库建设而言,它是建立在多年实施义务教育阶段大规模学业质量监测的基础上,并服务于改进与提高学生学业质量的深入拓展研究。

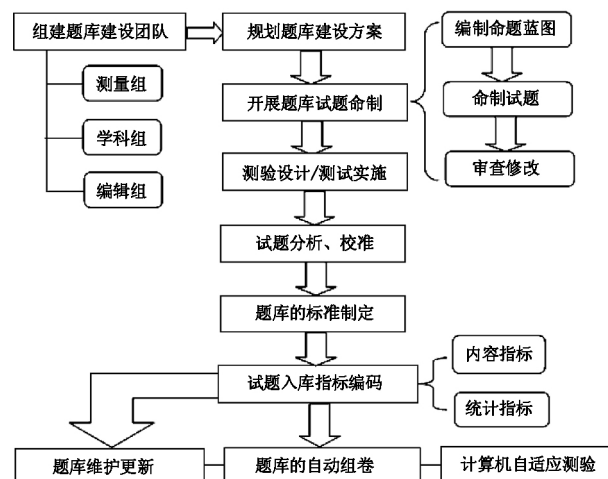


图1 题库建设流程图

在对国内外大量理论与实践研究进行文献分析的基础上,首先从专业团队的组建开始,整体规划题库建设方案,以教育目标分类学理论为指导,基于国家课程标准和测试方案命制试题,按照科学的测验设计实施测试;开展基于教育测量学理论的试题分

析、校准,即通过先验的试题难度和考生能力值将题库中所有试题放在同一量尺上;进行标准制定,即由专家小组按照严格的科学程序估计达到合格、良好和优秀水平的临界能力值;标定试题,即将试题按照既定的内容属性和统计属性指标遵循一定的规则顺序入库,在计算机程序算法的保证下,实现自动组卷;同时,做好题库维护更新工作;不断提高技术水平,最终形成计算机自适应测验(如图1所示)。

3.2 关键环节

3.2.1 组建题库建设团队

学业质量监测题库建设是一项需要花费大量人物财力、环节众多、程序复杂、周期较长的浩大工程,需要教育测量技术、学科命题技术与计算机编程技术的有机结合。因此,组建一支由教育测量、学科命题、计算机编程人员组成的高水平专业队伍是题库建设的首要环节。其中,教育测量组由具有教育测量学理论功底和大规模实测数据分析处理经验的专业人员组成,主要负责拟定题库整体建设方案、测验设计方案、数据分析方案、标准制定方案、指标入库方案、组卷方案等;学科命题组由具有学科教育教学理论背景和实践经验,掌握学科命题技术的学科专家、教研员和一线教师组成,主要负责制定学科测试方案、编制细目蓝图、命制试题等;计算机编程组由掌握计算机编程技术的专业人员组成,主要负责基于题库建设方案进行计算机程序开发设计,并形成最终的题库管理系统。

3.2.2 开展题库试题命制

学业质量监测题库试题的命制,主要包括依据国家课程标准制定测试方案、编制细目蓝图、命制试题和评分标准、审定试题等环节。

(1) 制定测试方案

学业质量监测题库测试方案,是对学科学业测试内容、结构、范围与标准的具体说明,是在对国际上的基础教育教学评价与学生学业评价体系,如美国的教育进步评价计划(NAEP)、国际经合组织的PISA项目等进行深入研究的基础上,由学科专家、教研员、教师组成的学科命题小组,依据国家课程标准相应学段内容与学科教学的实际,经过反复研究、讨论,广泛征求意见后最终形成,是指导学科测试工作的重要基础。一般包括学业测试框架、学业测试内容标准、测试方式与题型、测试领域的分数构成、学业成就水平等部分描述。其中,学科学业测试框架、学业测试内容标准与学生成就水平标准是构成

学科学业测试方案的三大核心要素。

(2) 编制细目蓝图

根据学业质量监测题库测试方案,编制学科命题细目蓝图,其主要目的是使试题在内容上具有代表性,在内容领域、难度、区分度分配上具有合理性,从而保证测验工具具有良好的结构。命题细目蓝图包括试题在内容领域、能力领域上的分布,试题描述及与内容标准的对应情况、错误选项的类型,以及试题预计难度、区分度等指标(张咏梅,田一,郝懿,2009)。

(3) 命制试题和评分标准

根据命题细目蓝图,组织专家分组进行平行试题及评分标准的命制工作。以国家课程标准为依据,遵循命题细目蓝图,结合具体试题类型(如客观题与主观题)的要求进行命制,形成初步的学业质量监测题库试题。此外,制定相应的评分标准也是命题的重要组成部分。其中,客观题的答案具有唯一性,评分标准制定略易。主观题的答案则较为开放,大多答案不唯一,评分标准制定相对较为复杂。

(4) 审定试题

审定试题是保证学业质量监测题库内容科学性、结构合理性的重要环节。为确保监测评价与课程标准、测试方案相匹配,借鉴美国Achieve“测验—标准”一致性分析工具(杨玉琴,张新宇,占小红,2011)构建适合于学业质量监测的一致性分析模式,即从测验试题符合国家课程标准的程度、兼顾不同版本教材内容的程度、区分不同学业成就水平的程度、各领域考查的层次比例合理性程度等方面对试题进行审定,以促进方案的完善和测验的改进。

3.2.3 题库等值设计与校准

纳入学业质量监测题库的所有试题,必须是建立在同一量尺上,否则试题之间无法进行科学的参数标定,因此需要进行科学的等值设计。题库的等值设计包括单一组设计(Single-group design)、相等组设计(Equivalent-group design)、定锚测验设计(Anchor-test design)、共同考生设计(Common-person design)等(余民宁,2009)。其中,定锚测验设计最为常用,它是将有一定数量共同试题的两份测验给予两组不同考生施测的测验设计方案,这些共同试题称之为锚题(Anchor Items)。该设计不仅实际操作性强,而且避免了其他设计所带来的测试时间长、考生疲劳或重复练习等无关因素的干扰。

在科学的等值设计方案基础上的试题分析与校

准是学业质量监测题库建设的核心环节,一般采用以项目反应理论为基础的等值方法进行试题数据校准,它包括在相应IRT模型上的试题参数估计,以及在等值设计基础上以先验试题难度和考生能力值为标尺将题库中所有试题放在同一量尺上的方法(Eggen & Norman, 2011)。目前常用的等值校准方法有平均数平均数法(Mean/Mean Method)、平均数标准差法(Mean/Sigma Method)、特征曲线法(Characteristic Curve Method)等。Cohen和Kim在1998年研究表明,上述校准方法等值结果相差不大(Cohen & Kim, 1998; Michael & Robert, 2004)。故学业质量监测题库的等值校准采用比较简明易懂的平均数标准差法。其基本原理是通过将不同测验上共同题的原始参数进行一定的函数转换,以确保获得相等的平均数和标准差的方法。假设某共同题在试卷A上难度值为 b_A ,在锚卷上难度值为 $b_{\text{锚}}$,建立两个函数关系 $\overline{b_{\text{锚}}} = \alpha \overline{b_A} + \beta$; $S_{\text{锚}} = \alpha S_A$,求得 α 和 β 的值,然后试卷A上的试题参数估计值,便可以转换到与锚卷相同的量尺上。

3.2.4 题库的标准制定

建设具有标准参照测验性质的学业质量监测题库时,还需要组织学科专家为试题库进行标准制定,即由专家小组根据国家课程标准和学科测试方案,结合实际测试的情况,按照科学程序估计达到合格、良好和优秀水平的临界能力值。目前标准制定方法包括Nedelsky法、Angoff法、Ebel法、Bookmark方法等。

现主要采用基于项目反应理论的Bookmark方法,该方法适于两级计分和多级计分两种题型,能够将多道测验试题放在同一试题手册中排序,从而达到设置合格、良好、优秀多重临界能力值的目的,既提高时间效率,又能降低临界能力值的标准误(陈梦竹、张敏强, 2009)。具体步骤为:将所有测试试题进行校准后按照IRT难度进行升序排列,然后邀请学科专家小组按照由易到难的顺序讨论每道试题,判断“刚刚合格的学生”是否能够“掌握”这道试题。“掌握”的界定一般为至少有67%的概率能够答对该试题(Karantonis & Stephen, 2006)。此后,在刚好能和刚好不能“掌握”的试题上作出标记,以这两道试题IRT难度的平均值作为合格能力值。对于良好、优秀能力值亦是如此制定。

3.2.5 试题入库指标编码

试题入库指标编码,是学业质量监测题库建设的重要环节。试题入库指标编码包括试题的内容属

性指标和统计参数指标两大类。

(1) 试题内容属性指标,是指对题库中试题的基本属性和内容结构的标定。基本属性主要是指试题入库的基本信息,包括试题编号、命题时间、命题人、修订时间、预测时间、预测对象、预测地区、修改时间、入库时间、使用次数、是否锚题等项目;内容结构是指有关学科试题的内容结构,包括题型、知识结构、标准答案(包括错误选项说明的问题)、领域等。例如,依据国家课程标准,数学学科题库中试题的知识结构需要在三个层次进行标定,如图2所示。



图2 题库知识结构层次图

(2) 试题统计参数指标,包括基于经典测量理论的统计参数指标(CTT指标)和基于项目反应理论的统计参数指标(IRT指标)。其中,CTT指标包括难度、区分度等。①难度,是描述试题难易程度的指标。两级计分试题采用通过率来说明难度,多级计分试题则采用该题的平均分与满分的商来表示。②区分度,指试题对考生某种特质的区分程度或鉴别能力,主要包括题总相关系数(R)和鉴别指数(D)两个指标。其中,题总相关系数采用将考生群体在某道试题的得分和相应总分计算相关系数的方式来求得,相关系数越大,区分度越高。鉴别指数分为合格鉴别指数和优秀鉴别指数,合格鉴别指数是将作答该试题的考生按照标准制定的合格线分为合格组 and 不合格组。两组在该试题上的得分率之差即为合格鉴别指数,两级间差值越大,合格鉴别指数越高,试题对于合格水平考生群体的区分度越高。优秀鉴别指数计算方法与相应内涵与此类同。

IRT指标主要包括三参数对数模型与等级反应模型的试题特征曲线、参数指标、信息量曲线等。①试题特征曲线(Item Characteristic Curve, ICC):采用单调增函数来反映考生潜在特质与正确作答概率的关系,此函数曲线确定考生对试题做出正确作答的概率随考生潜在特质水平的提高而增加。如图3所示。②参数指标:主要包括试题区分度(a)、难度(b)和猜测度(c)。在图3中,区分度指试题特征曲线拐点处的斜率,其值越大,说明试题对考生群体的区分程度越高。难度表示正确回答该试题概率为50%时所对应的考生群体的能力参数,也称为能力

值 θ 。猜测度是试题特征曲线的截距,其值越大,说明考生群体越容易猜对本道试题。③试题信息量曲线(Item Information Curve, IIC):反映试题能够为考试提供信息量的程度。好的试题对考生能力水平估计的误差较小,能提供较高的信息量。

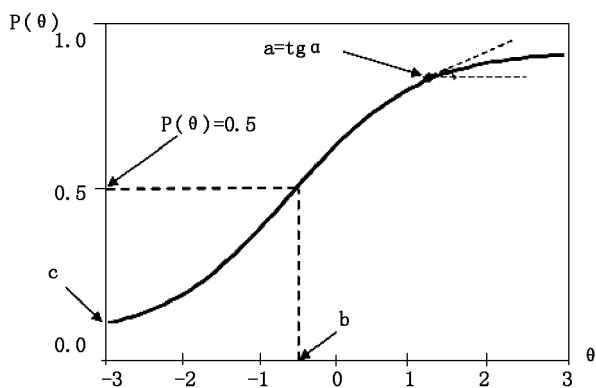


图3 基于项目反应理论三参数模型的试题特征曲线

3.2.6 题库的自动组卷

学业质量监测试题按照指标入库后,初级目标是通过计算机程序算法来实现自动组卷功能。学业质量监测题库的自动组卷如图4所示。

首先以学业质量监测题库测试方案的学科细目蓝图为基准,进行内容结构设置和 CTT 指标设置。内容结构设置:要保证组卷生成的试题类型、数量、分值、领域等内容结构与学科细目蓝图一致;CTT 指标设置:保证每道试题的 CTT 指标在预计值范围内,并且兼顾难度和区分度的范围界定值。根据前一项设置(方案1)或前两项设置(方案2),可以随机组出若干套试卷;

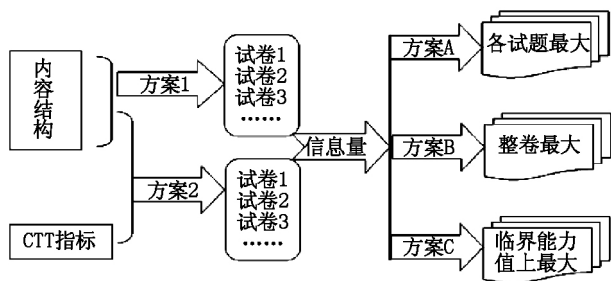


图4 题库的自动组卷示意图

基于上述设置结果并结合 IRT 指标,以信息量峰值最大为准则,再进行以下三种选择方案设计:方案 A:通过上述条件设置,每个设置点上可能有几个备选试题,在每个备选试题中选择信息量峰值最大的试题;方案 B:通过上述条件设置,组出若干套试卷;根据信息量曲线的分布及其峰值来决定选择哪一套试卷;方案 C:根据标准制定后所确定的合格、

良好、优秀学生临界能力值,选择在合格、良好、优秀三个临界能力值上信息量峰值最大的试卷。

3.2.7 计算机自适应测验

学业质量监测题库在实现自动组卷的基础上,正迈向实现计算机化自适应测验的进程。计算机自适应测验(Computerized Adaptive Test, CAT),是指测验试题自动显示在屏幕上,考生根据自己的判断按键作答。考生作答过程中,计算机按设计好的程序及时根据作答资料估出考生的可能反映水平,并针对这一水平迅速决策,从大型题库中调取难度恰当、性能优良(区分度大,猜测可能性小,从而信息量大)的试题继续施测,直到施测的试题足够多,测验信息量加和达到指定值为止。然后,计算机报告测验结果,包括分数和评语(林崇德等 2003)。

计算机自适应测验可以在任何时间任何地方对不同考生施测,适用于广泛的能力水平,既可以节省施测时间做到安全保密,又能够达到精确估计考生能力的目的。它以项目反应理论为理论依据,以大型题库建设为基础,通过测验信息量控制,使测量精度保持一致,保证了测验的精确性、公平性及实效性。美国教育研究协会等在 1999 年修订的《教育与心理测试标准》中,标准 3.12 指出“计算机自适应测验的理论依据和辅助凭证应记录备案。这份文件应包括用于选择单元试题的步骤、决定测验起始点和终止条件的步骤、测验评分的步骤和控制试题曝光的步骤”(美国教育研究协会等 2003)。

在计算机自适应测验的未来发展中,还需要进一步考虑结合认知诊断模型的认知诊断测验,以便能够更加精确有效地评估考生能力。

4 义务教育阶段学业质量监测题库建设有待进一步研究的问题

当前,义务教育阶段学业质量监测题库建设虽然取得了一些重要的研究进展,但仍然处于起步阶段,还有大量问题有待深入研究论证。其中主要包括:第一,研究开发学业质量监测题库的计算机自适应测验功能。在查阅大量文献并进行先期相应模拟研究的基础上,制定符合学业质量监测题库性质特点与需求的相应选题策略、估计方法、终止标准等,从技术上实现计算机自适应测验,从而加强施测的便捷性和保密性,增加测验的精确性、公平性和实效性。第二,研究开发基于学业质量监测题库的省或市级网上教学专家即时诊断反馈系统。在实现题库计算机自适应测验的基础上,根据不同学科年段课

程标准要求,不断补充、更新与教材相适应的试题,加强对学生学习和教师教学的即时在线诊断反馈,为学生学业质量提高和教师教育教学改进提供基于题库系统的支持。第三,加强关于题库建设的基础研究。基础研究是保证题库质量与科学性的根本,也是未来发展的必要条件。例如教育测量认知诊断技术的研究、主观题作答模式分类的研究、多种数据等值与试题校准方法的精度比较研究、多种组卷方案的比较研究、基于不同模型的计算机自适应选题策略等。

参考文献

- [美]布卢姆等编著.(1986).《教育目标分类学:第一分册认知领域》(罗黎辉等译).上海:华东师范大学出版社.
- 陈梦竹,张敏强.(2009).Bookmark法设置划界分数的研究述评.《心理科学进展》,17(5),1102-1108.
- 戴海琦,张锋,陈雪枫.(2007).《心理与教育测量》(修订本).暨南大学出版社.
- 林崇德,杨治良,黄希庭.(2003).《心理学大辞典》.上海教育出版社.
- 美国教育研究协会,美国心理学会,全美教育测量学会主编.(2003).《教育与心理测试标准》(燕妮琴,谢小庆译).沈阳出版社.
- 香港考试及评核局.(1977).取自 <http://www.bca.hkeaa.edu.hk/web/SA/zh/Introduction.html>.
- 余胜泉,朱凌云.(2000).《现代远程教育资源建设技术规范》(征求意见稿).[EB/OL].2000-11-08.取自 <http://www.edu.cn/html/keyanzf/doc/d.doc>.
- 余民宁.(2009).《试题反应理论IRT及其应用》.台湾:心理出版社.
- 杨玉琴,张新宇,占小红.(2011).美国Achieve“测验一标准”一致性分析工具的研究及启示.《教育科学》,27(4),86-90.
- 张咏梅,田一,郝懿.(2009).北京市义务教育阶段学业水平测验的开发.《教育科学研究》,9,37-40.
- Cohen, A. S. & Kim, S. H. (1998). An investigation of linking methods under the graded response model. *Applied Psychological Measurement* 22(2), 116-130.
- Eggen, T. J. H. M. & Norman, D. V. (2011). Item calibration in incomplete testing designs. *Psicológica* 32, 107-132.
- Karantonis, A. & Stephen, G. S. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice* 25(1), 4-12.
- Michael, J. K. & Robert, L. B. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.
- Wim J. Van der Linden, & Ronald K. Hambleton. (1997). *Handbook of modern item response theory*. New York: Springer Verlag.

Theory and Practice of Item Bank Construction for the Compulsory Education Academic Quality Monitoring

Tian Yi Zhang Yongmei

(Beijing Academy of Educational Sciences, Beijing 100191)

Abstract: Academic quality monitoring in the stage of compulsory education is an important way to evaluate students' academic quality and it is of vital importance in the current domestic and foreign education quality assessment. Therefore, enhancing the construction of item bank of academic quality monitoring is a necessary way to promote the organization and implement of the test tools from experience to science. Based on the elaboration of the properties and functions of the academic quality monitoring in the stage of compulsory education, the research discusses the basic theory, process and key steps of the academic quality monitoring item bank, and points out the problems to be further studied in the study of the current item bank.

Key words: compulsory education; academic quality monitoring; item bank construction