课题名称：基础教育学业评价与课程标准一致性分析模式的开发与应用
课题批准号：CBA14047
课题类别：青年专项
课题负责人：田一　副研究员　北京教育科学研究院
主要成员：黎坚　郝懿　李美娟　李英杰　崔红梅　王俊英　游晓锋

“基础教育学业评价与课程标准一致性分析模式的开发与应用”成果公报

“基础教育学业评价与课程标准一致性分析模式的开发与应用”课题组

一、研究背景

2001 年教育部正式启动新一轮基础教育课程改革，颁布《基础教育课程改革纲要（试行）》等一系列政策文件，经过十多年的改革发展，我国基础教育课程改革取得卓越成效。随着教育评价研究领域的蓬勃发展，基础教育学业成就评价成为了解课程改革实施质量的重要途径，各省市纷纷开发各种学业成就评价工具开展基础教育监测工作，但是区域评价工具的科学有效性如何保证成为基础教育学业成就评价的关键问题。而国家课程标准是基础教育学业成就评价的根本依据，因此，区域基础教育学业成就评价与课程标准的一致性分析的就显得尤为重要。所谓一致性分析，即分析、判断评价与课程标准之间的匹配吻合程度，属于效度研究范畴。

从国际范围看，20 世纪 80 年代，美国发起“由标准驱动并基于标准”的基础教育课程改革，并把课程与教学的一致性作为检测州、学校是否有效落实课程标准的一项关键性指标，各州必须证明其评价与课程标准具有很强的匹配程度。但是如何考查学业成就评价与课程标准之间的匹配程度成为基础教育课程领域研究的热点问题。1998 年美国成立课程与评价一致性分析协会，许多研究者和研究机构提出学业成就评价与课程标准一致性分析模式，其中以韦伯模式、Achieve 模式和 SEC 模式最具代表性。韦伯模式是 N. Webb 在 1997 年提出，主要以内容重点为核心，从内容领域一致性、知识深度一致性、知识广度一致性、分布平衡性等四个维度测查学业成就评价与课程标准内容要素的匹配程度；Achieve 模式是以 R.Rothman 的理论为基础，由美国非赢利教育研究机构 Achieve 公司罗伯特和思莱特主持组织开发的一致性研究的综合化工具。主要从向心性、挑战性、均衡性三个方面考察测验是否仅仅测量标准中所要求的内容、测验在多大程度上测量了标准中的核心内容以及测验对于学生是否具有足够的挑战性，系统分析了多方面的影响因素，构建了综合性较强的新型“学业评价-课程标准”一致性研究工具。具有浓重的韦伯工具的印迹，内涵接近。SEC 模式是美国在 20 世纪 90 年代基于课程标准的教育改革以及学校教学要求偏低的现实背景下，由威斯

康星州教育研究中心学者安德鲁•帕特和约翰•史密森以计划课程调查（Survey of Enacted Curriculum，简称 SEC）数据为基础共同开发和研制的评价与课程标准一致性水平分析程序和方法，强调一致性研究的整体化，主要用于进行课堂教学与学业评价之间的一致性比较，推动了美国学校基于课程标准的评价实践。此外，威克森模型、TIMSS 测验-课程匹配分析、斯坦福国际研究所 SRI 模式、2061 计划等也有效地弥补了各研究工具的实施方法。

从国内范围看，华东师范大学崔允漷团队对基于标准的学生学业成就评价的一系列研究（崔允漷，王少非，& 夏雪梅，2008；汪贤泽，2008），引起了国内学者对于学业成就评价与课程标准一致性的关注。当前研究主要集中在两个方面：一是美国经验，包括韦伯模式、SEC 一致性分析范式、Achieve 分析模式等的介绍（刘学智，& 马云鹏，2007；范立双，& 刘学智，2010；杨玉琴，张新宇，& 占小红，2011；岳喜腾，& 张雨强，2011）；二是基于美国一致性分析范式的学科应用研究，包括小学数学（刘学智，2008），小学语文（曹小旭，& 张庆霞，2011），小学体育（赵广涛，2010），高中化学（王后雄，孙建明，2013），高中物理（王焕霞，2012）等学科，主要是从学业水平测试质量分析的角度，对学科测试卷进行一致性分析，均暴露出学业水平考试内容与课程标准未能完全保持一致，目标领域下的知识掌握水平界定模糊，无具体学业标准可参照，教育命题技术欠缺等问题。此外，也有从区域学业成就评价与课程标准一致性分析的角度，开展小学数学学业水平测试与课程标准一致性水平研究，结果表明各地区编制的学业水平测试试卷在内容与结构上都存在着偏离课程标准的问题（刘学智，高云龙，2012）。

从北京范围看，北京市在 2002 年曾对各区县的初中毕业/升学考试评价工具进行效度研究；2009 年聘请 62 名各学科专家教师对人大附中等 10 所学校自主命制的 77 份高中会考试卷质量进行评价；2010-2013 年，北京市义务教育教学质量分析与评价反馈系统（Beijing Assessment of Educational Quality，BAEQ）聘请学科专家对其评价工具进行基于国家课程标准的一致性分析，作为该项目测验开发的重要环节。但总体而言，我国关于基础教育学业成就评价与课程标准一致性分析的实证研究还偏少，特别是当前区域基础教育学业成就评价对于标准的依据不够严谨，评价工具命制流程不够规范等现象，使得国际一致性分析模式的本土化应用有待研究。

基于以上分析，本研究依据国内基础教育改革和发展相关政策规划，充分借鉴国内外一致性分析研究经验，结合我国实际教育教学评价情况，探究本土化一致性分析模式，以分析区域基础教育学业评价与课程标准一致性情况。

二、内容与方法

（一）研究内容

本课题研究国际有影响的学业评价与课程标准一致性分析工具，主要借鉴韦伯和 Achieve 一致性分析模式，开发适合于基础教育学业评价与课程标准的一致性分析模式，并将其应用于语文、数学、英语等基础工具学科，分析其学业评价与课程标准的一致性水平，探索影响基础教育学业评价与课程标准一致性水平的因素，并对基础教育学业评价及命题提出相应的建议。

（二）研究方法

主要采用量化研究与质性研究相结合的技术路线，主要包括文献研究法、访谈法、专家评判法。①文献研究法是研究基础性工作，并贯穿于整个研究过程，查阅途径包括使用网上数据库或图书馆藏书籍。②访谈法通过访员和受访人面对面地交谈，以口头形式根据受访者的答复搜集客观的、不带偏见的事实材料，以准确地说明样本所要代表的总体的一种方式。③专家评价法，即聘请教育评价领域专家与学科教学评价专家依据国家课程标准等相关文件要求及命制测试与调查工具的技术要求对相关测评材料进行评价。具体步骤如下：

1. 查阅文献，开展前期基础性研究工作。以查阅以往研究文献和报告为主要的手段，查阅途径包括使用网上数据库或图书馆藏书籍。主要查阅：①关于基础教育阶段国家课程标准相关文献；②关于学业标准制定与开发文献；③关于测验效度研究相关文献；④国际有影响的学业评价与课程标准一致性分析工具（例如韦伯模式、SEC 模式、Achieve 模式等）；⑤国内关于学业评价与课程标准一致性评价探索研究文献。

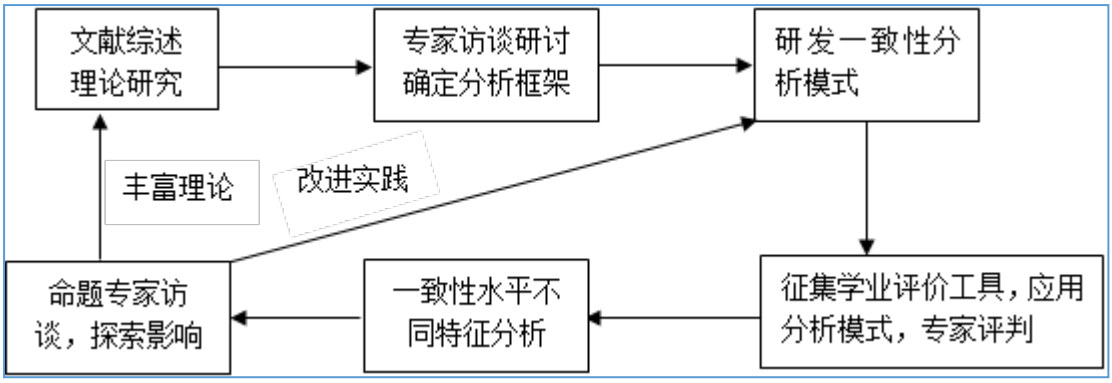
2. 邀请课标组专家、测量评价专家以及学科命题专家，从课程标准、测量信效度和命题技术等角度进行访谈研讨；

3. 基于专家访谈编码结果，同时结合前期文献研究及国际一致性分析模式，构建基础教育学业评价与课程标准一致性分析框架和维度，开发相应的评价工具。

4. 征集北京市各区县语文、数学、英语学科的学业水平测验工具，聘请教育评价领域专家与学科教学评价专家依据国家课程标准等相关文件要求及命制测试与调查工具的技术要求对区县测评材料进行评价。

5. 基于专家评价分数，进行数据统计分析，一是分析不同学科学业评价与课程标准一致水平的总体特征；二是分析不同区县各学科学业评价与课程标准一致水平特征；三是分析不同年级各学科学业评价与课程标准一致水平特征等。

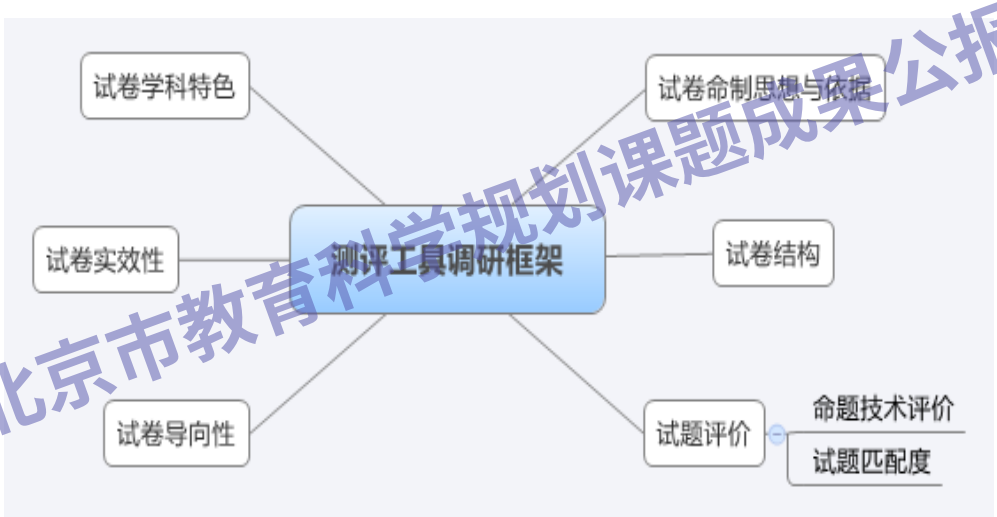
6. 在分析一致性水平特征的基础上，与专业命题者进行访谈，从命题者的命题取向角度、命题者对课程要素认识角度、命题者专业水平角度，探索影响基础教育学业评价与课程标准一致性水平的相关因素。



三、结论与对策

（一）一致性分析框架

在 BAEQ 评价工具内容效度量表的基础上，借鉴国内外一致性分析模式，依据相关国内外研究与已有的研究实践，构建本研究的评价框架——区域基础教育学业成就评价与课程标准一致性分析框架，如下图所示。



（二）编制分析工具

基于研究框架编制研究工具，包括试卷内容评定问卷和试题内容评定表（见附录）。其中前者从试卷命制思想与依据（7 道题目）、试卷结构（7 道题目）、试题评价（11 道题目）、试卷导向性（7 道题目）、试卷实效性（2 道题目）、试卷学科特色（5-9 道题目）等方面进行调查，采用李克特量表的形式，从完全不符合到完全符合五点计分；后者主要评定与课程标准内容目标、能力目标的匹配程度，与细目蓝图内容目标、能力目标的匹配程度等，从完全不匹配到完全匹配五点计分。经教育测量学指标分析，该研究工具内部一致性信度达到 0.97，各维度与总分之间的相关系数均在 0.91-0.93 之间，具有很高的信效度。

（三）主要结论

1. 纵观区域基础教育学业成就评价工具，试题评价维度表现最好，体现了一定的命题

思想。

通过对北京市 2012-2013 年度区域基础教育学业评价与课程标准一致性研究，从专家试卷内容评定问卷的数据来看，各区县在试题评价维度得分最高，表明各区域均具有一定的命题经验和技巧，在具体题目的命制方面体现了专业性和科学性。例如图 8 中的小学数学试题，考查“图形与几何”领域的内容知识点，其中前两道题目属于基础知识的理解和掌握，第三道题目要求学生基于给定的条件自己设计方案，是在基础知识上的灵活运用，特别是结合创设情境解决实际问题，对于学生问题解决等高级思维能力有所测查，体现了一定的开放性和灵活性。此外，部分学科借鉴国际大型测验项目 PISA 的测试理念，结合实际生活经验，考察学生的非连续性文本阅读，恰恰考查学生的这种实际生活阅读能力，与国际近年来的测试理念相符，有助于提高学生实际生活的能力。

2. 区域学业成就评价与课程标准内容目标的匹配程度较高，而与能力目标的匹配程度较低。且部分区域没有细目蓝图，无法考查其匹配程度。

从专家试题评定表的数据来看，大部分区域学业成就评价与课程标准内容目标的匹配程度均在 4.5 左右，而与能力目标的匹配程度普遍较低。这可能是由于部分学科课程标准中的内容目标比较具体，具有一定的可操作性，而能力目标方面偏宏观，笼统不具体，可操作性不强，使命题评价的时候没有针对性。同时，结合学业成就评价测试工具来看，有些试题还是存在超过课程标准要求或课程标准未涉及的内容，对青少年认知思维发展规律重视不足，死记硬背题目偏多，缺乏问题解决等高级思维能力方面的考查。

此外，研究也发现部分区域没有相应的评价框架和细目蓝图，或者细目蓝图编制较为简易不科学，无法体现试卷命制依据和思想，这一方面使测试工具与细目蓝图的匹配程度无法考量，另一方面也使试卷内容评定在试卷命制依据和思想维度上得分偏低。但同时我们也可以看到，对于有细目蓝图的区域而言，其学科试题与细目蓝图匹配程度较高，表明各区县均能很好地基于具有实际可操作性的目标进行评价工具编制，展现很强的操作实践能力。因此，将课程标准的内容目标和能力目标具体化，更好地编制细目蓝图作为命题依据，是开展学业成就评价的核心基础。

3. 区域学业成就评价工具还存在一些科学性错误，评价形式过于单一。

结合对各区县评价工具和专家建议的文本分析来看，就评价工具而言，区域学业成就评价工具过多采用北京市监测原题或相似题等；部分试题存在较多的科学性错误，这将给学生带来极大困惑，也给科学评价带来困难。此外，还有部分试卷知识点考查覆盖单一，同一知识点反复测查。例如小学科学测试卷对于“摆的快慢与长度”这一知识点在填空题、判断题

和选择题三种题型中反复考查，一方面占用卷面内容考查同一知识点，造成了评价资源的浪费；另一方面该试卷对本知识点掌握好的学生有利，而对其他学生不利，使学业成就评价分析有失公平。

就评价形式而言，区域学业成就评价大多采用的是传统纸笔测查，评价形式过于单一，这也是导致评价工具在试题导向性和学科特色维度得分偏低的原因。根据国际大型测验项目的经验，采用多元化评价形式是科学考查学生能力的有效途径之一。除传统纸笔测验外，还有人机交互计算机测验、口语交际测验、现场操作测验等实践型测验，而基于该测验形式的表现性评定是当前国内外研究的热点问题。它在测查学生的高级思维能力和综合运用所学知识解决实际问题的能力，激发学生的学习动机以及优化教学过程方面有显著作用，同时也为当前的考试和评价改革提供了新的思路。因此，区域基础教育学业成就评价如何在纸笔测验的基础上引入实践测验，开展表现性评定，是以后区域教育评价实践的重点研究问题。

4. 区域小学数学学业成就评价与课程标准一致性最好；各学科均表现为年级越高，其与课程标准的一致性程度越高。

从区域各学科学业成就评价与课程标准一致性的分析研究来看，小学数学学科在试卷内容各维度评定指标和与课程标准匹配程度上均得分较高，其次为小学语文、英语、科学，小学品德与社会学科的得分偏低。这一方面由于小学数学学科课程标准清晰可操作程度较高，各区域在命制评价工具时参照性很强；另一方面也可能与学科教师的师资力量、课时量的保证、学科评价的熟练程度等因素有关，小学数学、语文、英语等学科在区域学业成就评价中均处于优势地位，而科学、品德与社会等学科受重视程度不高，加之课程标准的具体化待加强，使其一致性表现较差。

从年级角度来看，六年级在试卷内容各维度评定指标和与课程标准匹配程度上得分最高，其余年级学业成就评价与课程标准的一致性程度逐渐降低。这可能与高年级学生在知识内容考查和评价方面越来越严格规范化，而低年级学生在学业成就评价方面的要求偏低有关。因此，必须重视对于低年级学生学业成就评价与课程标准一致性的关系，保证各年级学生均获得高质量的学业成就，推动基础教育课程改革的顺利进行。

（四）对策与建议

（一）认真学习国家课程标准，充分利用市级评价框架，结合实际学业标准，开展学业成就评价工具编制工作

国家课程标准是义务教育教学质量评价的基础和核心，贯彻执行国家课程标准要求是保证义务教育均衡公平发展的根本。因此，建议各区县教研部门和学校教师深入理解课程标准，

明确其中的具体内涵，注重课程标准与实际生活的联系，渗透到评价工具的编制中。同时，北京市级层面组织课程领域、学科领域、测量领域专家，花大力气构建市级测试框架。通过研究发现目前因各区县研究水平良莠不齐，建议能尽量采用市级测试框架（因市级框架集合了更多的专家团队进行精心打造，具有很强的标准性），做到资源的充分利用。

目前市级层面已经并且完成了部分学科的学业标准开发，针对当前各区县基于标准命题的不明确性，以及框架蓝图等环节的缺失，建议市级层面继续将模糊不清晰的课程标准，结合北京实际情况，开发各年级的学业标准，建立统一标准，开展市级培训解读，使各区县各学科教研部门、一线教师能更加明确课程标准，为更好地指导教师教学、学生学习服务。

（二）规范命题流程环节，推行审校制度，保证命题把关的有效性

经过区域学业成就评价与课程标准一致性的研究，结合专家命题经验，发现各区县在命题环节存在着很大差异，有些区县命题环节比较完备，有些区县仍然停留在凑题评测阶段。因此，建议各区县在以后的评价工具编制中，要保证命题工作环节的完整性，使各区县命题流程更加完善、科学、规范。例如成立命题小组、明确职责，认真负责，增强试卷规范化，杜绝科学性错误，使试题的描述更加规范准确。

对于命题制度和流程不规范的区县，建议推行审校制度，要有评价工具的预测环节、编制细目蓝图，规范图表及语言表述，要重基础，宽覆盖，难易适中，杜绝科学性错误和低级错误。

此外一定要保证命题把关的有效性。对命题质量的追踪评价、审查、专家评定、一致性分析、效度研究等。做到命制的试题从科学角度信效度要好；应用角度贴近学生实际，符合课程标准要求，遵循学生思维发展规律；使评价工作更加科学有效，对教师的教和学生和学反馈指导更具针对性。

（三）把握命题原则，做好命题技术培训，提高教师命题能力

命题技术是整个评价工具开发编制的核心技术环节，做好命题的根本就是掌握科学有效的命题技术。

针对本次研究过程中发现的各区县学科命题技术方面的问题，建议各区县教研部门和学校教师把握命题原则，研究命题技术；建议市级教研部门做好教师命题技术培训工作，提高教师命题能力。例如，细目蓝图的编制——依据框架，结合试卷命制的依据和思想，将内容标准、知识点明确的渗透在细目蓝图中，为命制试题打下基础；“蓝图”中应包括知识与能力考查的双向细目（内容、题型、认知层次）及分值分配比例。又如主观题、客观题、评分标准的命制规则和方法等。

（四）采用多元化评价方式，加强实践类项目测查，同时关注高级思维能力的培养和考查

除传统纸笔测查之外，开展多方式化考查。特别是对于实践类项目的考查方式，例如语文口语、英语听说、科学类实验操作等，建议采用多元化测查手段，使用面对面交流、人机对话、实际操作和现场演示的方式，采用表现性评定方法，全面科学地了解掌握学生能力发展状况。同时，当前各区县基于国家课程标准的评价，主要还停留在基本的认知领域层面，对学生是否记住知识点、简单的应用和实践方面有所测查。但是对于高级思维能力测查方面有待进一步加强。因此建议在常规命题环节的基础上，注意学生高级思维能力的测查，例如问题解决能力，目前 PISA2012 年开始测查和分析学生具体情境的问题解决能力，这是高于学科之上的一种综合能力的测查，这一方面有待市级教研部门和区县教研部门、一线教师共同研究探讨，更多的关注学生不同思维水平、认知加工方式的评价与教学。

总体而言，北京市各区域基础教育学业成就评价，要依据国家课程标准，注重课程标准与实际生活的关联；树立发展性评价观；遵循青少年认知思维发展规律；规范命题流程，加强命题技术培训，注重实践测查，关注高级思维能力发展，使评价工具的编制更加科学严谨，使对于学生和教师的评价更加客观有针对性，最终促进北京市义务教育教学质量的全面提高和发展。

四、成果与影响

（一）课题研究成果

1. 一致性分析模式及工具

在前期文献研究基础上，邀请课标组专家、测量评价专家以及学科命题专家，从课程标准、测量信效度和命题技术等角度研发基础教育学业评价与课程标准的一致性分析模式，并编制相应的工具，包括试卷内容评定问卷和试题内容评定表。

2. 研究报告

基于基础教育学业评价与课程标准一致性分析模式及工具实施，对北京市 17 个区县小学语文、数学、英语、品社、科学等学科开展一致性分析，并撰写《小学教学质量工具调查研究报告（总层面）》、《小学教学质量工具调查研究报告（语文学科）》、《小学教学质量工具调查研究报告（数学学科）》、《小学教学质量工具调查研究报告（英语学科）》、《小学教学质量工具调查研究报告（科学学科）》、《小学教学质量工具调查研究报告（品社学科）》及东城区、西城区等 17 个区县和地区的《小学教学质量工具调查研究报告（XX 区）》，共计 23 份研究报告。

3. 期刊论文

课题负责人以第一作者在国家教育类中文核心期刊上发表文章 3 篇，为《学业质量监测与课程标准一致性研究》刊登于《上海教育科研》（2016 年 9 月），《义务教育结果公平现状及趋势的实证研究》刊登于《教育科学研究》（2016 年 10 月），《区域基础教育学业评价与课程标准一致性的本土化研究_以北京市为例》刊登于《教育测量与评价》（2016 年 10 月）。

（二）成果影响

本课题研究的基础教育学业评价与课程标准一致性模式及相应的研究工具，充分应用于北京市小学教学质量监测工具的调研中，通过专家的定性和定量的评判，分析出各学科及各区县编制该学科学业评价工具的现状及存在的问题，并结合相应的问题提出有针对的改进意见。所反馈的 1 份市级总层面报告、5 份学科层面报告和 17 份区县层面报告为全市学业评价工具研发、学科命题及区县测评等环节提供了丰富的实证依据。最终发表的论文成果多次在院学术年会和国内教育测量年会上做分享报告，在国内教育评价领域形成了一定的影响力。

五、应用与采纳

课题研究的一致性分析模式及工具（包括试卷内容评定问卷和试题内容评定表）连续多年被北京市义务教育教学质量分析与评价反馈系统作为审题环节的重要工具，为专家从定量角度对学业评价命题提供科学抓手。

课题所撰写的总层面报告为市级教育行政部门和教研部门提供了丰富的实证依据，并为其所采纳，应用于教育教研指导工作；所撰写的分学科报告为各学科教学专家提供了实践支撑，为学科专家指导各区域命题提出了有针对性的科学建议；所撰写的各区域报告为各区县教育行政部门和教研部门提供实证资料，并为区县考虑教研命题的集中科学提供新的视角。

课题最终发表的论文《学业质量监测与课程标准一致性研究》被下载 444 次，被引 2 次；，《义务教育结果公平现状及趋势的实证研究》被下载 361 次，被引 4 次；《区域基础教育学业评价与课程标准一致性的本土化研究_以北京市为例》被下载 263 次，被引 3 次。

六、创新与改进

1. 基于国际成熟的一致性分析评价模式，结合国内基础教育教学情况，构建适用于国内基础教育学业评价与课程标准一致性分析的框架和维度，如何使该框架符合国内基础教育现状，维度涵盖国内基础教育课程标准一致性分析的内容。本研究充分调研国内基础教育教学领域专家和骨干教师，研发出一套适合我国基础教育课程标准一致性分析的框架和模式。未来还可以拓展到课程、教材等评价载体与课程标准的一致性研究。

2. 开发基础教育学业评价与课程标准一致性分析模式的研究工具，如何使该研究工具

能够科学有效全面地反映一致性程度。本研究在前期文献和访谈研究基础上，课题团队基于一致性分析模式研发测评工具，从试卷命制思想与依据、试卷结构、试题评价、试卷导向性、试卷实效性、试卷学科特色等方面编制试卷内容评定问卷，同时基于与课程标准内容目标、能力目标的匹配程度，与细目蓝图内容目标、能力目标的匹配程度编制试题内容评定表。但对于更加细致化的基于内容标准的逐条评价分析还有待完善。

3. 如何做好一致性分析模式的分析与应用，是该模式应用于基础教育一致性分析评价领域的落脚点，也是检验该分析模式本土化是否成功和有效的根本。本研究在研发完成一致性分析模式并编制研究工具后，在北京市 17 个区县进行本土化验证，并将结果应用于各区县学业质量评价的命题指导工作中，具有很强的实践意义。而将指导意见更好地应用于学科命题工作中将是未来实践的重点。