我国大规模教育评价项目探究与实践

王　蕾

[摘　要]国际上三个比较有影响力的基础教育评价项目在评价理念、技术、手段和结果报告等方面都有值得我们借鉴之处。教育部考试中心开展PISA2006中国试测研究在管理标准、技术标准和数据标准上吸取了PISA的优点。我国开展科学、有效的大规模教育评价项目，使教育评价成为促进教育发展和提高教育质量的有效途径，应整合大规模教育评价研究与实施的平台，增强评价的学术性，大力培养专业人才，利用信息技术实现跨越式发展。

[关键词]教育评价；大规模教育评价项目；PISA；TIMSS；NAEP

[中图分类号]G40-058.1　[文献标识码]A　[文章编号]1009-7187(2007)11-0025-04

如何开展有效的教育评价，使教育评价成为促进教育发展和提高教育质量的有效途径，目前在我国还缺少比较系统的理论研究和大量成功的科学实践，所以关注国外大规模基础教育评价项目的评价理念、技术、手段、结果报告及其发展动向，特别是通过实践深入学习并熟知国外大规模基础教育评价项目的整个评价流程体系，掌握国际教育评价的设计理念和操作方法，为我所用，无疑将大大推动我国基础教育评价领域的研究和发展。本文通过对当前国际上三个比较有影响力的基础教育评价项目——学生能力国际评价（PISA）、国际数学和科学趋势研究（TIMSS）及美国国家教育进展评估（NAEP）的比较研究，结合教育部考试中心开展PISA2006中国试测研究的实践经验，阐述对我国如何开展科学、有效的大规模教育评价项目的思考和建议。

一、PISA、TIMSS和NAEP简介

学生能力国际评价（Programme for International Student Assessment，PISA）是经济合作与发展组织（The Organization for Economic Co-operation and Development，OECD）发起并组织实施的，为参与国家和地区协作监控教育成效的评价项目，测试主要工业化国家义务教育阶段结束后15岁的学生在阅读、数学和科学方面所具备的应用知识和技能解决问题的能力。PISA通过收集学生的背景信息，进行包括个人、家庭和学校等方面的因素解构，形成“学习成果质量、公平性和均衡分布、学习者的特征、学习风气、学校资源和学校政策与实践”[1]等教育成效评价指标体系，为参与国家和地区政策分析和研究提供有价值的参考。PISA于2000年首次开始启动，每三年进行一次，周而复始。PISA2000、PISA2003、PISA2006分别有32个国家、41个国家和地区、56个国家和地区参与。[2]为了保证评价的效度和信度，来自各参与国家和地区的教育政策制定者和相关领域的专家共同决定评价的范围、本质、学生背景信息收集等，评价材料也考虑到不同的文化和语言影响，其翻译、取样和资料收集过程都采取严格的质量监控机制，并通过实施大规模实地预试等各种手段，将测试在各参与国家和地区实施中可能存在的误差降到最小。

国际数学和科学趋势研究（Trends in International Mathematics and Science Study，TIMSS）由国际教育成就评价协会（International Association for the Evaluation of Educational Achievement，IEA）发起，其主要目的在于为各参与国家和地区制定课程和教学政策提供建议，提高各国和地区数学和科学的教学水平。TIMSS还对学生、教师及校长进行问

王　蕾　教育部考试中心　100084

卷调查，考察学生数学和科学的学习背景，以阐明家庭和学校对学生成绩的影响。TIMSS 于 1995 年发起，每四年为一个周期，评价目标为四年级和八年级学生的数学和科学成绩的发展趋势。参与 TIMSS2003 的 48 个国家和地区既包括发达国家也包括发展中国家。[3]

美国国家教育进展评估（National Assessment of Educational Progress，NAEP）主要由国家评估管理委员会（National Assessment Governing Board，NAGB）和教育部下属的国家教育统计中心（National Center for Education Statistics，NCES）共同完成。国家评估管理委员会为 NAEP 制定政策，编制评价框架和测试规范；国家教育统计中心负责执行。NAEP 从 1969 年起定期实施，测评美国学生在各个主要学科领域的知识和技能，为教育者和政策制定者提供当前美国学生成就水平的最新状况，并基于之前的评估，比较分析得出学生成就的变化趋势。NAEP 以四年级、八年级和十二年级的学生为测评对象，主要评估学校课程和国家课程共同包括的知识和技能，即特定的内容主题和广泛的思考技能，其评估领域涵盖阅读、数学、科学、写作、美国历史、公民以及地理，最主要的是阅读、数学和科学；[4] 还有根据国家和各州需要所设计的评价。在四年级和八年级的阅读、写作、数学及科学评价中，参与各州可对有代表性的样本进行建构，将评价结果与州目标相比较，与其他州或全国的学生平均水平相比较。

二、PISA、TIMSS 和 NAEP 比较

美国国家教育进展评估项目（NAEP）是美国为自己量身定做的。NAEP 利用国家建立的监测框架（例如：每个学科的成就水平划分为基础的、精通的、高级的），搜集国家教育重点阶段各学科成就方面的信息。其他两个国际项目是由参与国家和地区共同决定的，以国际比较为重点。PISA 更强调对教育系统成效以及对“素养”的评价，其他两个评价项目的目标更侧重于课程设置的成效。

通过对 PISA、TIMSS 和 NAEP 的比较，我们可以看出，这些国外基础教育学生成就表现评价项目在评价理念、内容、目的、对象取样和结果报告上存在一些异同。

（一）评价理念上，每一个评估体系都是在其哲学思想和框架下进行的。TIMSS 的理念是评估学生在基本知识和概念方面，与课程框架紧密联系的数学和科学学科的思维能力。PISA 的理念是评估学生在日常生活情境中处理问题所需要的阅读、科学和数学素养。NAEP 的理念是了解不同年级学生对不同学科的深入理解情况，以便于最佳配置美国学生所需要的知识、技巧和能力。

（二）评价内容上，TIMSS 和 NAEP 大多与学校课程有密切的联系，试图测量学生对具体知识、技能和概念的掌握程度，大量题目覆盖课程的内容，评估单一的、确定的某一知识、概念和能力的掌握，只有少数题目测试学生科学和数学的综合思维能力。而 PISA 则侧重于测量广义的“素养”。[5] PISA 测量的素养是 15 岁的在校学生，为迎接当今不断变化的现实世界的挑战，应用知识和技能解决问题的能力，以及在日常生活情境下做出良好判断和决策的能力。它不同于且高于对学校课程所设置的学科相关知识的理解或记忆能力的考察。

（三）评价目的上，TIMSS 和 NAEP 主要指向于学校，所收集的学生背景信息侧重于反映不同国家的教师是如何进行教学的，以及这些教学对学生的成绩可能产生的影响。而 PISA 则体现了更强的为国家教育决策服务的目的。PISA 的目的在于衡量各国义务教育的产出，应用测试结果提供教育成效对比的有效指标，并通过收集家庭、学校等方面的相关信息，为国家教育政策制定和调整提供参考。

（四）评价对象取样上，PISA、TIMSS 和 NAEP 都采用抽样进行评价。TIMSS 和 NAEP 采用以年级为基础进行抽样的方式，主要报告课程成绩。PISA 采用以测试时年龄在 15 岁 3 个月到 16 岁 2 个月的学生为基础进行抽样的方式，目的是描述义务教育结束时的教育成效。

（五）评价结果报告上，PISA 以 30 个 OECD 成员国的平均值为基准[6]。OECD 成员国大多是西方主要资本主义国家，教育质量国际平均值代表了当今世界发达国家的平均水平。TIMSS 的参与国家和地区即包括工业化国家，也包括全世界的中等收入和发展中国家和地区，公布的国际平均值以所有参与国家和地区为基准。NAEP 的参与国为美国，报告反映美国学生当前的成就水平。

三、PISA 在我国的实践和启示

PISA 注重测试学生未来发展的潜能，关注对学生的人文素养、知识运用、探究能力和情感态度的考察，其题目以现实生活为背景，超越了对课程具体内容的知识考核，这对于解决当前我国教育评价改革中的难题有很大的借鉴意义，对于国内基础教

育多元化发展，分省命制中考题，2007年新高考方案实施后全国基础教育评价工作如何展开有一定的启示意义。

PISA可以对世界发达国家的教育发展状况做出比较全面的评价。我国基础教育成效世界瞩目，若参与PISA，可以通过对比评价和监控我国教育体制的效力与发展。教育部考试中心2006年引进并启动了PISA2006中国试测研究项目。PISA2006中国试测研究并不代表我国正式参与PISA，目的在于学习、借鉴PISA先进的评价理念和技术，了解国际的情况，通过实践锻炼队伍，构建符合中国国情的评价标准、手段、技术和方法体系；促进考试内容和形式的改革，特别是对命题环节的改进，有利于全面推进素质教育。

教育部考试中心PISA2006中国试测研究实施了PISA测试工具翻译和预试调整、学校样本和学生样本选取、评价实施、编码阅卷、数据整理、统计分析和结果报告全环节的评价工作。

（一）管理标准上，PISA2006中国试测研究在操作层面上实行两级管理体系，即教育部考试中心负责研究工作在全国的实施管理，试点机构在统一要求下负责本地区的组织实施，包括本地区抽样信息提供、学校样本和学生样本落实、人员培训、测试及调查问卷实施管理，并要求对测试的题册和数据严格保密。教育部考试中心统一编码评判试题册和问卷，统一录入信息，保证编码评分和数据录入的信度和效度。数据分析由教育部考试中心自主完成。通过PISA2006中国试测研究，教育部考试中心不但完成了试点地区的教育成效评价报告，还利用PISA协作组织提供的国际参数进行国际比较，为各级教育决策者提供了一个多层次评价教育成效的指标体系和国际评价参考信息。

（二）技术标准上，PISA2006中国试测研究按PISA国际规范采用两阶段分层随机抽样设计，第一阶段完成学校层面的抽样，选取了五个分层变量：学段，包括初中、高中；学校的地理位置，包括市区、县城（县政府所在城镇）和农村；学校的性质，包括公立学校、私立学校；学校的类型，包括普通中学、职业学校、特殊教育学校和中等专业学校；学校的等级，包括普通高中、示范高中、普通初中、基础薄弱初中。试测研究按照这五个维度设计了抽样框架，试点地区据此框架上报了本地区所有包含15岁学生学校的统计信息。教育部考试中心在第一阶段抽取了150所样本学校，随后在第二阶段以完全随机抽样的原则从这150所学校样本中每所学校抽出35名15岁学生作为参加测试的学生样本。共有来自试点地区150所中学的5000余名学生被纳入样本，样本有效地代表了试点地区近1200所学校的16万余名15岁在校学生总体，其中农村学校在校学生占将近一半。

PISA2006中国试测研究采用的是纸笔测验，共有13套试题册，依据随机原则将每名学生分配到每套试题册，每个学生需用两个小时的时间完成测验。PISA2006中国试测研究收集的原始数据同时采用国际大型统计软件SPSS和SAS两套系统进行数据清理和转换，得到完全匹配的结果，保证了数据的精准。随后，用清理后的数据按照OECD数据分析的标准流程对每个学生样本依据所属不同学校、所做不同试题册、学生问卷收集到的不同背景类型进行回归分析，每个评价领域生成5个似真值（plausible value，PVs）[7]，对PVs进行加权，通过抽样权重的多重复制程序（replicate）[8]进行分析，按国际规范得到了评价结果。评价结果可直接与参与PISA2006的56个国家和地区相比。PISA2006中国试测研究数据分析保障了试题抽样和学生抽样数据分析的最大程度似真，在时间短、学生样本量小、试题覆盖面广、成本低、误差小、减轻学生负担、减少考试焦虑的同时，确保了评价的科学性，拓宽了评价的内容与形式，使教育评价更具实用价值。

（三）数据标准上，PISA2006中国试测研究的数据可实现与国际横向的比较，奠定了未来纵向跟踪研究的数据库基础。PISA2006中国试测研究对数据进行科学和全面的深入分析，多角度、多层次地提供分析报告，从定量实证研究的角度评价教育成效。PISA2006中国试测研究依据数据分析结果形成了试点地区平均、各试点地区、学校样本和学生个体样本四个层面的数据分析报告，为国家教育整体质量的提高，为地方教育质量的改善，为学校、学生提供了服务。

参与PISA试测研究，对于我国教育评价相关领域的研究与实践具有十分重要的启示与借鉴作用。PISA评价理念先进科学，评价标准公正规范，评价程序周密严谨，评价结果全面可靠。借助PISA2006中国试测研究结果，我们首先可以了解我国义务教育阶段结束时青少年的能力表现情况，并且通过与其他国家同年龄学生进行比较，从中发现我国教育制度的优势与不足，借鉴在PISA评价中表现优秀国家的经验，为我国的基础教育体制改革提供参考意

见并提出合理建议；其次，通过分析 PISA 的评价结果，我们可以从中发现影响学生成就的诸多因素，如学校因素、教学因素及动机因素，并对此提出改进意见，这对于提高我国的教育质量具有重大意义；第三，鉴于目前我国的教育评价体系相对于国际水平仍十分薄弱的现状，PISA 测试流程的整体实施，使我国相关教育考试机构学习并熟知了国际教育评价的设计理念和操作方法，有助于建立满足我国教育当前需求的教育评价体系，在全国范围内开展科学有效的大规模教育评价项目。当然，在借鉴当前国际先进的教育评价理念的同时，我们需结合当前我国的教育现状，探索适合我国国情的教育评价理论和方法。总之，PISA2006 在我国试测研究的成功无疑将大大推动我国大规模教育评价项目的发展，并为后续研究提供基础。

四、对我国开展大规模教育评价的思考和建议

我国的基础教育已进入全面提高教育质量阶段。如何开展科学、有效的大规模教育评价项目，使全国性教育评价成为提高学生素质的有效手段，成为实现教育管理和教育决策科学化的重要保障，笔者有如下思考和建议。

（一）整合我国大规模教育评价研究与实施平台

随着对于大规模教育评价重要性认识的提高，我国许多机构正在利用各自的资源优势开展相关研究。北京师范大学正在筹建“中小学义务教育质量监测中心”，中央教育科学研究所正在开展全国教育科学“十一五”规划国家重点课题“中小学生学业成就调查研究”，人民教育出版社正在进行“中小学生学科学业评价标准的研究与开发”科研项目。国内各种大规模教育评价虽然在评价理念、目的和作用上正如 PISA、TIMSS 和 NAEP 各有侧重，但在一定程度上制约了我国当前大规模教育评价投入产出功效最大化，不利于整体协调发展。PISA、TIMSS 和 NAEP 都由美国国家教育统计中心协调执行，三个项目的评价报告从结果和技术上可以相互验证比较，为美国社会提供美国教育状况的全方位分析与预测。国内大规模教育评价应走集约化发展的道路。

（二）增强我国大规模教育评价的学术性

大规模教育评价本身是门学问，评价的理念、目的和作用、技术和手段以及结果报告等在国际上都有很大的学术提升空间。我国大规模教育评价要用教育测量与评价的最新理论来指导，使评价的功能得到充分发挥。从以上大规模评价项目的经验来看，加强教育测量专家、学科专家和数理统计专家的结合，发挥各自所长，是提高我国大规模教育评价质量和效果的关键因素。

（三）大力培养我国大规模教育评价专业人才

大规模教育评价是一种专业化水平很高的工作，在 PISA2006 中国试测研究实践中遇到的最突出的困难是缺乏抽样、教育测量、学科教学及数据分析等专业人才。大规模教育评价在我国不能再停留在纸上谈兵的阶段，而能够胜任大规模教育评价工作的实际操作人员无论在专业素质上还是在数量上，都无法适应和满足我国日益深化的教育评价改革与发展的迫切要求。PISA2006 中国试测研究提供了国际先进水准的大规模教育评价案例，同时提供了可快速有效地培养出既了解现代教育测量与评价理论，又掌握大规模教育评价各环节操作的高素质教育评价专业人才的途径。

（四）利用信息技术实现我国大规模教育评价跨越式发展

信息技术和大规模教育评价实践的结合是当前教育评价发展的重要趋势，更是大规模教育评价成功实现的必要保障。PISA 在应用网络平台实现全球指挥管理，计算机自适应测试研究，抽样和数据分析专业化软件开发，网上电子评价报告自动生成等方面都为信息技术在大规模教育评价中的实际应用树立了榜样。计算机自适应测试结合因特网更可以提供超越时空的弹性施测环境，节省相当大的社会实施成本，尤其在我国幅员辽阔，教育信息建设蓬勃发展的状况下，更可以发挥远距离评价的优势和实用性。在我国的大规模教育评价中引入计算机自适应测验无疑会推进我国教育评价与国际的接轨，实现大规模教育评价跨越式发展。

[注释]

[1] OECD. PISA2003 country profiles [EB/OL]. http://www.pisa.oecd.org/document/50/0,3343,en_32252351_32236173_37627442_1_1_1_1,00.html, 2006-12-21.
[2] OECD. Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006 [M]. Paris: OECD, 2006.8-44.
[3] IEA. Countries participating in TIMSS2003 [EB/OL]. http://isc.bc.edu/timss2003i/countries.html, 2007-6-25.
[4] NCES. The nation's report card [EB/OL]. http://nces.ed.gov/nationsreportcard,2007-6-29.
[5] OECD. Learning for Tomorrow's World: First Results from PISA2003[M]. Paris: OECD, 2004.23.
[6] OECD. PISA2003 Technical Report [M]. Paris: OECD, 2005.185-216.
[7][8] OECD. PISA2003 Data Analysis Manual [M]. Paris: OECD, 2005.31-80.

（责任编辑：韩 梅）