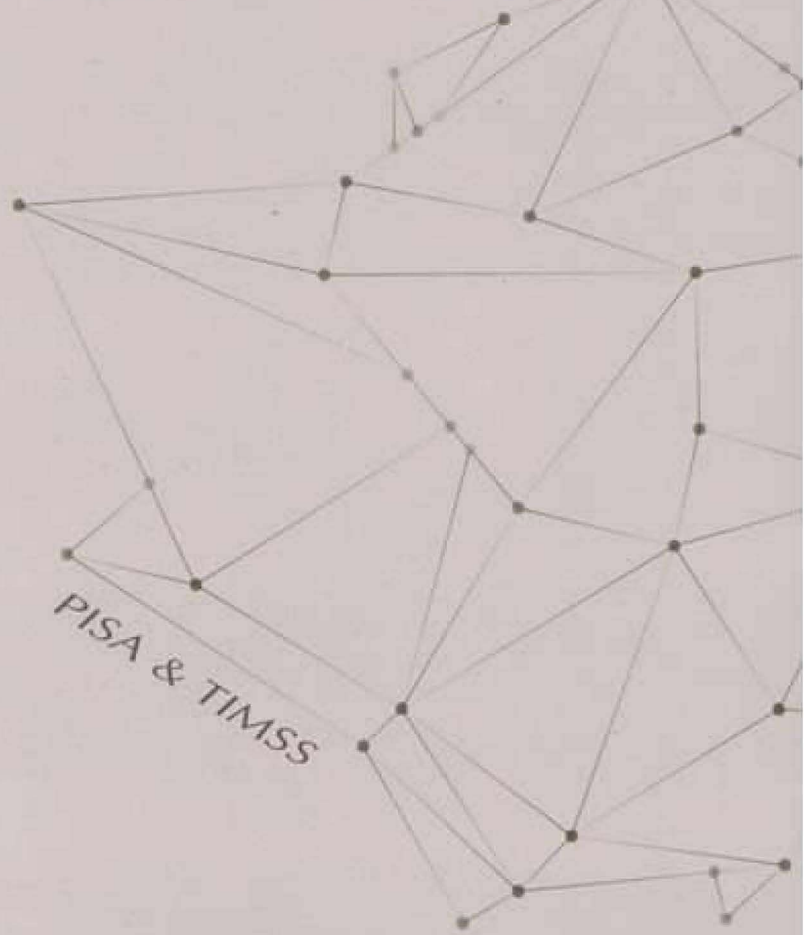
国家社会科学基金资助项目

国际视野下大规模数学测评研究

王鼎 著



上海科技教育出版社

图书在版编目(CIP)数据

国际视野下大规模数学测评研究/王鼎著. —上海:上海科技教育出版社,2017.9

ISBN 978-7-5428-6607-3

Ⅰ.①国… Ⅱ.①王… Ⅲ.①数学教学—教育评估—研究 Ⅳ.①01-47

中国版本图书馆 CIP 数据核字(2017)第 210365 号

责任编辑 卢 源

封面设计 李梦雪

国际视野下大规模数学测评研究

王 鼎 著

出版发行 上海科技教育出版社有限公司

(上海市柳州路 218 号 邮政编码 200235)

网 址 www.sste.com www.ewen.co

经 销 各地新华书店

印 刷 上海师范大学印刷厂

开 本 787×1092 1/16

印 张 18.75

版 次 2017 年 9 月第 1 版

印 次 2017 年 9 月第 1 次印刷

书 号 ISBN 978-7-5428-6607-3/O·946

定 价 49.80 元

序 PREFACE

目前，中国正在推进教育综合改革，考试招生制度改革是其中的重中之重。提升考试评价的专业化水平，是考试招生制度改革的必由之路。考试评价是一项系统性、综合性很强的工作：系统性体现在考试测评系统内相关要素的功能及要素间的相互联系上；综合性体现在多种方法和技术的综合使用及整合上。同时，系统性和综合性需要做到有效整合。考试评价涉及课程设置、课堂实践和学生学习成就等因素，蕴含着课程实施、学生学习和考试评价之间的一致性，甚至还联系着平衡和协调教育资源、促进教育公平和提升教育质量等诸多问题，因此对考试评价的设计及组织实施过程有着更高的要求。如何通过考试评价系统性和综合性的整合，确保考试评价的公平性和考试结果的可靠性、有效性，是当下在考试评价领域需要作出回答的重要问题。

中国正在深入推进新一轮的课程改革。学生核心素养的提出，以及学科核心素养框架的建立，呼唤着考试评价的改革。如何科学、合理地把握好考试评价这个杠杆，撬动课堂教学和课程实施的有效性，为学生的学习进步和全面发展，为教学的改善和提升提供足够的证据，是值得深入思考的问题，也是很有意义的挑战。王鼎博士的专著正是在此背景下，尝试对上述问题和挑战作出积极回应。

从 19 世纪末至今，教育考试评价的历史发展经历了四个时代，分别是测量时代、描述时代、判断时代和心理建构过程时代。现在有学者开始尝试构建第五代教育评价理论。第五代教育评价理论侧重于对评价组织、目标、价值和结果的重新解释与构建，其核心是评价结果的修正，以及如何反馈教学实践。可见，考试评价已从反映客观现象的大致描述，逐步转向对人的思考过程的精确定位。这一发展趋势一方面反映了人类对于自身认知的深入研究，以及心理测量技术的快速发展，另一方面更凸显出上述两者的有机融合。上述变化正在影响考试评价的理念、设计及相关技术运用，乃至结果解释。因此，我们有必要对考试评价的本质进行重新认识，并随之调整相应的设计。在这方面，国际大规模测评有很多成熟的经验可资借鉴，如 PISA 测评和 TIMSS 测评中的多维度、多层次、多种工具、多种目标的系统实证方法等。

随着中国考试招生制度改革的逐步推进，建立科学、公平、可靠、有效的大规模测评变得越发重要。本书以数学学科测评为切入点，通过对 PISA 和 TIMSS 数学测评的比较分析，归纳出

国际大规模数学测评的系统架构及特点，并借助量化实证分析方法，关注中国大规模数学测评在框架设计、命题设计、结果解释等方面的问题，关注相关差异对测评结果的影响，以及学习机会等因素对测评结果的影响等，思考和分析中国大规模数学测评的系统化设计及专业化实践，旨在为大规模数学测评的建设及改进提供建议。我相信，本书的出版一定会对中国的考试招生制度改革起到积极的推动作用。

夏惠贤

2017 年 6 月 28 日

内容摘要 ABSTRACT

20 世纪中后期，随着对人力资本在社会经济发展中作用的不断认识和愈加注重，各国或地区日益关注本国或本地区教育对于人的培养，关注教育过程及教育体系的质量和效果。作为现代科技发展中必备的基础性和工具性学科，数学成为现代经济发展中人才的基本素养之一。数学测评在人的素养甄别及发展性需求上，正逐步受到重视。鉴于测评结果在课程实施及政策层面的重要性，数学测评在整个系统构建及结果解释上日益受到重视。

本书将聚焦中国义务教育阶段结束时的大规模数学测评——初中数学学业水平考试。在中国，初中学业水平考试承担着对义务教育阶段进行学业水平评定的任务，甚至还承担着为高中阶段学校进行选拔的任务。由于它的高利害性，从命题到最后的成绩报告，以及相应的组织管理，都需要极其谨慎的态度和精确细致的工作。

在中国，对学科测评的标准和规范的研究越来越受到关注。如何有效准确地进行学科测评，无论是在社会性层面，还是在测评技术层面，都是大家极其关注的问题。这显然不能简单等同于命制试卷，而是需要在宏观层面上系统地建立相应的学科测评体系。但在这方面，无论是理论设计，还是实践操作，系统性的研究并不多见。而且，目前国内大规模数学测评在系统设计、命题理念及测量技术运用上存在的难点，越来越影响着测评结果的有效性及结果的解释力。本研究结合国际大规模数学测评如 TIMSS 和 PISA 数学测评的相关理论及经验，结合具体案例及相应量化模型，深入实证分析中国的大规模数学测评，对其下一步的发展进行思考，并提出相应策略。

基于上述考虑，本书分八章进行论述。在相应章节中呈现如下内容。

第一章是绪论。对目前中国大规模数学测评研究的背景、研究意义进行简要叙述，再就本书中要涉及的相关概念进行扼要说明。

第二章将论述重心转移到国际，就国际大规模数学测评进行简单叙述。主要涉及的内容包括：国际大规模测评发展的驱动因素、国际大规模数学测评的现状，以及针对国际大规模数学测评的国内外研究情况。其中国际大规模数学测评的相关内容，主要以 TIMSS 和 PISA 数学测评为代表。这两个测评也是目前国际上最大、最有影响力的数学测评项目。

第三章是本书后几章论述的理论储备。内容主要针对以 TIMSS 和 PISA 数学测评为代表

的国际大规模数学测评系统，并以给中国大规模数学测评系统构建提供借鉴和解决思路为目的，确定以认知、观察和解释为测评系统的基本维度，建立以这三者为基本要素的大规模测评系统比较分析框架。

第四章分别对 TIMSS 和 PISA 数学测评在不同要素上的历年变化进行纵向梳理，第五章则对 TIMSS 和 PISA 数学测评的不同要素进行横向比较。通过这种纵横的分析比较，进一步明确各个要素在国际大规模数学测评中的定位、表现及相互间的内在联系，明确国际大规模数学测评系统的发展特征，以及其综合性、内在一致性、连续性的特点，明确其建立所需的基础。

第六章简要叙述目前国内大规模数学测评的现状。内容主要涉及测评背景，如课程改革及学科教育目标等，也涉及中国大规模数学测评系统设计的现状及其难点。

第七章在第六章的基础上，为考虑建立适合中国的大规模数学测评系统，进一步从量化的角度，对测评系统建设进行解析，其中既涉及测评系统本身的结构维度设计，又涉及测评结果的影响因素模型。结果显示，在测评框架设计的不同维度上，无论是内容分布还是过程分布的差异，都会对测评结果产生直接影响。此外，数学学习结果影响因素的影响显著性程度也存在差异。这就要求我们必须建立适合自己国家的大规模数学测评系统，不能盲目照搬国外现有的测评系统。对于不熟悉量化模型的读者而言，可以直接阅读本章的相应结论。

最后的第八章就构建适合中国大规模数学测评系统所要注意或着重突破的方面作进一步归纳和总结。该章明确了为保证大规模数学测评系统的内在一致性等要求，在测评目标框架的构建、共同核心内容与核心能力的择取、问题解决框架的制定等方面需要进一步开展的工作。

期望本书论述的内容能够有助于提升中国大规模数学测评系统建设及测评结果的解释力，为国内数学课程的实施及评价改革提供依据和参考，也为其他学科的大规模测评建设提供借鉴。

目录

CONTENTS

第一章　绪　　论

第一节　中国大规模数学测评研究的背景

上海代表中国连续参加了 2009 年和 2012 年的两轮国际学生评估项目(Program for International Student Assessment，简称 PISA)测评，这是由经济合作与发展组织(Organization for Economic Co-operation and Development，简称 OECD)发起的针对 15 岁孩子的国际测评项目。之后，上海与北京、广东、江苏一起参加了 2015 年的 PISA 测评。上海在前两轮的表现可以用“惊艳”一词来形容，在各个测试领域的成绩，均在所有参加测试的国家和地区中位列第一。

笔者选取既参加了 PISA2012 数学测评，又参加了中考数学测评的九年级学生根据这些学生的 PISA 数学成绩和中考数学成绩，绘制成散点图，如图 1－1 所示。

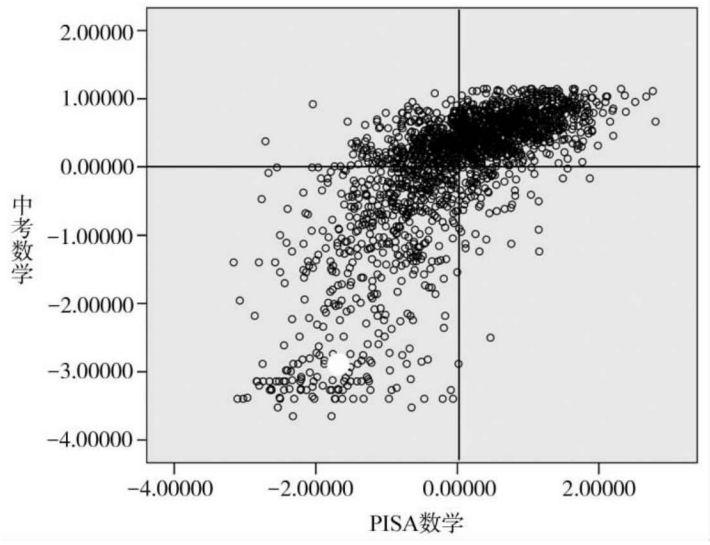


图 1－1　2012 年某市九年级学生 PISA 数学成绩和中考数学成绩散点图

上述散点图中，PISA 数学成绩采用了标准化的 logistic 值，中考数学成绩采用了标准化

的结果，两者的相关性是 0.714。散点图显示，几乎所有的群体都集中在第一、二、三象限。第一象限显示，中考数学成绩位于顶端的学生，在 PISA 数学成绩上呈现出一定的离散性。第二象限显示，PISA 数学成绩中等偏低的学生，在中考数学成绩上可以表现出中等偏高。第三象限显示，中考数学成绩低的部分学生，在 PISA 数学成绩上不一定低。同一个群体在两个大规模测评上呈现出不同的表现特征，这促使我们去思考两个测评之间的差异性。对这种差异的思考和分析不应仅仅局限于两者的测评内容，而更应关注两者在整个测评系统层面的区别。

这种比较，在一定程度上能为中国大规模数学测评的设计和实施提供更多实质上的参照，即可在数学测评系统层面探究相关大规模测评内在构建的基本要素及相关联系。最终目的是，以上述研究所得的共性和个性的特征作为基点，审视中国大规模数学测评（包括中考数学测评）领域的相关工作，期望对中国大规模测评工作有所帮助。

中国大陆如火如荼的课程改革及考试招生制度改革正在不断深入，测评与课程改革的一致性是人们日益关注的焦点。从测评本身的角度来看，为保证测评过程中各项技术运用的综合性，以及测评结果的可靠、有效性，对测评系统的设计是值得重新审视的。如果这些都做不到，就更不用谈测评与课程改革的匹配性问题了。

总体而言，本书对中国大规模数学测评的研究，主要是在三个背景之下进行的，分别是中国对 PISA 数学测评的引进、人的数学素养的重要意义，以及教育测评面临的新挑战。

一、中国对 PISA 数学测评的引进

PISA 测评在中国的施测，使得我们对国际大规模测评项目从“远观”走向了“近赏”。国家教育部考试中心从 2006 年开始实施 PISA 中国试测研究项目，进行了 PISA 测评工具翻译和预试调整、学校样本和学生样本提取、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工作。上海从 2008 年开始，参加 PISA 测评相关工作，分别在 2009 年和 2012 年参加了 PISA 正式测评，在阅卷、数学、科学三个领域，都取得了第一名的好成绩，在华夏大地上引起很大震动。PISA 测评也逐渐从象牙塔走进了人们的日常言谈之中。同时，由 PISA 测评所引起的反思也不断呈现在我们眼前。它测了什么？是否真的有效？对我们的教育有何启发？这类问题不断呈现在一些相关文章中。①

对于中国测量或评价领域而言，PISA 测评是新鲜事物。面对其所得测评结果数据，可以解释多角度多层面的问题，甚至可以为宏观教育政策提供数据支撑。人们不免会思考：为何中

① 这些文章包括：

孔凡哲，李清，史宁中.PISA 对中国中小学考试评价与质量监控的启示[J].外国教育研究，2005，05：72—76.

傅禄建，沈祖芸.PISA2009：国际视野中的上海义务教育质量　访上海市教育委员会主任薛明扬[J].上海教育，2010(24)：22—23.

罗阳佳.PISA 带给我们什么[J].上海教育，2010(24)：16—18.

徐斌艳.关注第一背后的潜在不足[J].上海教育，2010(24)：41.

国的大规模考试(特别是高利害性考试)很难做到这些?还有许多人常为出现"科学的考试,不简单;简单的考试,不科学"现象而困惑。我们确实应该认真全面地分析和思考类似于 PISA 测评这类国际大规模测评所呈现出的"图景"。

国际数学和科学评测趋势(the Trends in International Mathematics and Science Study,简称 TIMSS)测评是比 PISA 测评出现更早,且影响同样巨大的国际比较研究项目之一。两个项目有一些共同的测评领域,数学就是其中的一个。同样都是测评数学,这两个国际测评项目的调查结果一直存在着差异,众多文件资料都注意到这一点。① 同样都是国际大规模的测评,整个测评管理及技术规范水平都非常高,却仍会得到不同的结果。是什么影响了这些结果的差异性?如何影响的?是否还有类似的结果存在?这些又说明了什么呢?在这些结果的形成过程中,测评系统为它们的有效呈现提供了怎样的保障和支撑?内在的因果逻辑是否存在?

二、人的数学素养的重要意义

随着信息技术的迅速发展,数学正不断渗透到其他学科。特别是 20 世纪中叶以来,数学与计算机技术的结合在许多方面直接为社会创造了价值,也使得数学在社会生产生活中的运用越来越广泛和深入,更加凸显出它在当今社会中的重要性和影响力。基于对数学重要性的重新认识,欧洲成立了"欧洲工业数学联合会",以加强数学与工业的联系,同时培养工业数学家去满足工业对数学的要求。学者沃尔伯格(H. Walberg)说:"从世界各国对学校学习内容评价的兴趣上,可以看到科技素养和经济产品的必然联系,而其中数学是首要的、被广泛认同的现代科技语言"。②

从这个意义上来说,数学已经成为社会经济生活中有着重要影响的学科,也是各国基础教育中非常重要的一门课程,它对提高每一个民族的科学和文化素质起着非常重要的作用。与此同时,人们对个人数学素养的培养,也提出了更新、更高的要求。在新的《义务教育数学课程标准(2011 年版)》中,开宗明义地指出:"数学作为对于客观现象概括而逐渐形成的科学语言与工具,不仅是自然科学和技术科学的基础,而且在人文科学与社会科学中发挥着越来越大的作用。……数学是人类文化的重要组成部分,数学素养是现代社会每一个公民应该具备的基本素养。作为

---

① 资料包括:

AIR. Reassessing US international mathematics performance: New findings from the 2003 TIMSS and PISA[M]. Washing, DC: American Institutes for Research, 2005.

Brow, G., Micklewright, J., Schnepf, S., & Waldmann, R.. International survey of educational achievement: How robust are the findings[J]. Journal of the Royal Statistical Society: Series A. 2007, 170(3): 623—646.

Grønmo, L., Olsen, R.. TIMSS versus PISA: The case of pure and applied mathematics. Retrieved February 2008. http://www.timss.no/publications/IRC2006_Gronmo&Olsen.pdf. 2015 - 12 - 6.

Hutchison, G., & Schagen, I.. Comparisons between PISA and TIMSS—Are we the man with two watches? in T. Loveless(ED.), Lessons learned—What international assessments tell us about math achievement[M]. Washington, D C: The Brookings Institution, 2007.

② Walberg, H.. Scientific literacy and economic productivity in international perspective[EB/OL]. http:// www.jstor.org /stable/pdf/20024851.pdf? acceptTC=true. 2016 - 3 - 2.

促进学生全面发展教育的重要组成部分，数学教育既要使学生掌握现代生活和学习中所需要的数学知识与技能，更要发挥数学在培养人的思维能力和创新能力方面的不可替代的作用。”①

三、教育测评面临的新挑战

在21世纪，人类面临种种更加严峻的挑战，各国之间的竞争更加激烈，比以往任何时期都需要更多的、更全面发展的人，从而对人的素质提出了更新、更高的要求，越来越需要更富创造性、更有适应性、更具个性化、更能够适应未来的全面发展的人。

除了需要更健全的心理和主动适应变化的品质外，我们还需要更全面的知识和能力。一是需要扎实宽厚的基础知识和基本技能。基础知识和基本技能相对较稳定，适应性较强。二是需要合理的知识结构。当前自然科学与社会科学相互渗透，科学与技术日益密切结合，知识日趋综合化，这就要求人们的知识结构应以本专业为中心，兼通文理，科学与技术相结合，理论与应用相结合。三是需要多样的能力。随着职业转换的加快，跨专业领域的活动日益频繁，国际化趋势的增强等等，单一的专业或职业能力已不能适应社会的需求，必须集自学能力、职业能力、创造能力、人际交往和合作能力、国际交往能力等各种能力于一身，才能更好地适应未来。

针对各国在各个领域的人才竞争日益激烈和对人的全面发展需求不断扩大的现状，教育领域已成为各国政府实现人的全面发展的“主战场”和发力点。为了达到目的，教育的质量和公平应成为全民素质教育培养目标的主要保障和主流价值，这已得到各国的广泛认同。教育部原部长袁贵仁在《深化教育领域综合改革》中也提出：深化教育领域综合改革必须有利于促进公平、提高质量。

随着客观条件的日益丰富和成熟，世界各国的教育日益关注实现人的全面发展的目标，在课程设置、课程实施、课程评价等方面作出了相应的变革。基于上述对于数学在现代社会中重要作用的充分认识，在各国的基础教育中，数学均是最重要、被安排最多教学时间的课程之一。各国均在各自教育系统中的不同水平层次上，对数学教育进行重新调整和梳理。以2001年6月出台的《基础教育课程改革纲要(试行)》为标志，中国的新一轮基础教育课程改革全面启动。随后，陆续出台了2001年《义务教育阶段国家数学课程标准》、2011年《义务教育数学课程标准(2011年版)》。上海在1998年启动上海中小学第二期课程教材改革(以下简称上海二期课改)，陆续出台《面向21世纪中小学课程方案和各学科教育改革行动纲领(研究报告)》(1999年)及《上海市中小学数学课程标准(试行稿)》(2004年)。这些新的纲领或课程标准，无论在课程的理念，还是在课程目标及目标体系、相应学科内容等方面，与原有的相比，都作出了调整。②

中国在义务教育阶段数学课程的理念上，关注学生的全面发展，倡导面向全体学生，适应学

① 中华人民共和国教育部制定.义务教育数学课程标准(2011年版)[M].北京：北京师范大学出版社，2011.1.

② 孔企平.上海数学课程：挑战、改革与反思.课程范式的转换——上海与香港的课程改革[C].上海：上海科技教育出版社，2004：99.

生个性发展的需要，课程内容反映社会的需要、数学的特点，符合学生的认知规律。国家《义务教育数学课程标准(2011版)》在总目标下设计了知识技能、数学思考、问题解决和情感态度四个方面；《上海市中小学数学课程标准(试行稿)》(2004年)在总目标下设计了知识与技能、过程与方法、情感态度价值观三个方面。总体而言，新的课程标准在关注学生全面发展的同时，在学科上除了原有数学知识和技能以外，还凸显了对于数学思维过程及方法的重视，注重解决数学问题的过程中对于问题的提出和问题解决结果的反馈分析，在原有基础上逐步强化了数学的应用。

教育测评系统建构过程中，在系统设计、内容设置、指标控制、结果呈现等方面如何体现课程理念及要求的变化，促进人的全面发展，以及如何体现教育公平，日益成为一个迫在眉睫的重大课题。

在中国，课程改革至今，从测评角度来说，应更好、更有力地体现出课程的设计理念，以促进课程改革的进一步深入。测评正成为提升课程改革发展的重要环节。上海市教育科学研究院胡兴宏研究员等在《关于上海二期课程教材改革的研究》中指出，现行的考试评价方式是二期课改深入进行的瓶颈和矛盾焦点所在。这个研究报告提到，“相当多的校长和教师认为，课程改革理念好，也是一个很好的方向，难以推广的关键还不是文件研制和教材编写中的不足，而是评价考试体系没有改。在座谈中校长们将此表述为：考试不改，再好的设想也是空想。”[①] 在回答“您领导学校实施二期课改过程中，遇到最大的困难是什么”时，校长们选择最多的选项是：课程理念与现有的考试评价方式不一致，所占比例为40.2%。在该研究报告中同时指出，政府要承担起减少现行考试评价制度负面效应和支持学校教改的责任，并提出与新课程相配套的评价体系有待建立。总之，该研究报告明确指出，由于方式的单一及结果运用的不合理，当前大规模考试作为教育实践中评价学生、教师、学校等的主要手段，对于课堂教学未起到积极的反拨作用。这值得引起社会各方的关注。

第二节 中国大规模数学测评研究的意义

本书对于中国大规模数学测评的研究，主要聚焦于初中数学学业水平考试(简称中考数学测评)，且不直接从国内的测评现状着手，而是将PISA和TIMSS数学测评作为突破口，先研究国际大规模数学测评的相关内容，再以此为鉴，反思中国大规模数学测评。

现有国际大规模数学测评，其目的在于进行国际间的数学教育比较。通过测评，为各国确立一个可供比较的公共平台，让各个国家对自身数学教育系统的优劣高低进行评估并探寻原因。PISA数学测评和TIMSS数学测评是目前国际上影响最为广泛的两个国际大规模数学

① “上海二期课程教材改革研究”课题组.关于上海二期课程教材改革的研究[J].上海教育科研，2005(05)：36.

测评。它们为我们提供了国际大规模数学测评的典型案例，成为我们把握国际大规模数学测评的内在结构、系统功能及特点的一个视角或分析框架。

鉴于上述，本书研究的目的是通过对 PISA 测评和 TIMSS 测评的纵向发展和横向归纳、比较及分析，梳理并整合国际大规模数学测评系统在基本目标框架设计、试题表征呈现、结果解释等方面的共同性、差异性、表现特征、构成要素及形成原因，尝试对数学测评系统这一整体的内部构成机理进行探讨。同时，本书尝试将中国大规模数学测评置于国际大规模数学测评的平台上进行研究，将上述国际大规模数学测评在基本目标框架设计、试题表征呈现、结果解释及系统构成机理等方面的特点予以归纳和总结，并对中国大规模数学测评进行反思，将定性分析与量化实证分析的方法相结合，力图为中国初中大规模学业水平考试数学测评系统的改善提供一个可行的操作依据，也为新一轮考试改革中相关内容及能力构建等提供好的参考和借鉴。

从这个角度出发，本书主要探讨以下这些问题。国际大规模数学测评具有哪些基本特征和构成要素？基于何种原因或思考形成了该国际大规模数学测评？该测评的内容为何如此选择？该测评的结果如何起到效果？在测评系统的结果呈现上，能为中国数学中考系统建设及相应改革带来怎样的启示和有益的支撑？

具体而言，根据美国国家研究理事会（National Research Council，简称 NRC）的“测评三角”（具体内容在本书第三章中有较为详细的论述，这里为了进一步明确问题，先提出相关概念），本书分解成如下具体问题。

问题一：PISA 数学测评和 TIMSS 数学测评形成的背景或原因是什么？

问题二：PISA 数学测评和 TIMSS 数学测评系统是在不断发展的，在这发展过程中，各个测评系统在目标维度设计、试题表征呈现、结果解释上作了哪些相应调整或变化？为何这样变化？上述三者内在的联系在各个测评系统中是否存在？它们是如何维持的？上述各个测评系统中的变化和不变说明了什么？我们可以从中获得何种启示或规律性的把握？

问题三：PISA 数学测评系统和 TIMSS 数学测评系统在发展过程中是否存在趋同性？如果存在，有哪些具体方面？体现出哪些相似的特征？我们如何看待这些相似性？这些相似性或差异性对结果呈现有什么影响？两个测评系统间存在的相似和差异又能进一步告诉我们什么？

问题四：结合中国大规模数学测评，特别是中考数学测评的现状，以 PISA 数学测评和 TIMSS 数学测评为代表的国际大规模数学测评对中国的中考改革及测评体系的完善有哪些启示或借鉴？

综上所述，本书自始自终的研究落脚点在于将中国大规模数学测评置于国际大规模数学测评的平台上，即在比较、归纳、总结国际大规模数学测评的相似性、差异性及其形成原因的基础上，将中考数学测评作为研究的主攻方向。这个做法将国际大规模数学测评的目标设计理念、对数学价值的理解、试题的表征及有效技术方法相结合，它所形成的相应结果解释有利于我们把

握学生在数学学业成就评价上的发展方向，为我们进一步推进数学测评系统的改进和考试改革提供一个可行的操作依据，也为我们新一轮考试改革中的内容改革提供有益参考和借鉴。

本书旨在建立并拓宽中国中考数学及中国大规模数学测评与国际大规模数学测评之间的沟通桥梁，为改进和完善中国中考数学测评提供有力支撑。为便于上述目的的达成，便于对PISA和TIMSS数学测评的比较分析，本书仅针对该两大测评的笔试部分进行研究分析。

第三节 相关概念分析

一、测评和国际大规模数学测评

说到测评(assessment)，首先要说一下测量(measurement)和评价(evaluation)。在著名测量专家史蒂文斯(S.S.Stevens)看来，测量是根据法则给事物赋予数量，即“用一定规则给事物指派数值或符号的过程”。[①] 教育测量，从广义上而言，是依据一定的法则(标准)，用数值来描述教育领域内事物的属性，是事实判断的过程。[②] 数学教育测量，是依据一定的原理和法则，用数值来描述数学教育领域内的事物属性(数学教学效果和学生的数学知识、数学能力)，并进行实施判断的过程。它是用一定的量尺来提供量化资料，从数量上来表现数学教育现象，具体来讲可以涉及学业、兴趣、适应性、智能等数学教育和心理方面的现象。[③]

评价“是一种价值判断的活动，是对客体满足主体需要程度的判断。”[④]评价的本质是价值判断，并借此收集信息，提供决策，完善工作，以此实现相应价值的过程。教育评价的概念是由美国教育家泰勒(R.Tyler)在其主持的“八年研究”中正式提出的，他认为评价在本质上是一个确定课程与教学计划实际达到教育目标程度的过程。而统计专家克隆巴赫(L.J.Cronbach)则把教育评价广义地定义为，为获取教育活动的决策资料，对参与教育活动的各个部分的状态、技能、成果等情报进行收集、整理和提供的过程。评价专家斯塔弗尔比姆(D.L.Stufflebeam)在1969年就直接提出评价是为决策提供有用信息的过程的看法。这也对评价实践产生了深远的影响。结合学科特点，数学教育评价可以界定为“全面收集和处理数学课程与教学的设计与实施过程中的信息，从而作出价值判断、改进教育决策的过程。”[⑤]

从上述表述分析可知，测量以评价为目的，评价以测量为手段，教育测量为教育评价提供

① 张敏强.教育测量学[M].北京：人民教育出版社，1998.4.
② 金娣，王刚.教育评价与测量[M].北京：教育科学出版社，2007.10.
③ 马云鹏，孔凡哲，张春莉.数学教育测量与评价[M].北京：北京师范大学出版社，2012.2.
④ 陈玉琨.教育评价学[M].北京：人民教育出版社，1999.7.
⑤ 马云鹏，张春莉.数学教育评价[M].北京：高等教育出版社，2003：15.

依据，是教育评价信息的主要来源。同时，教育测量结果只有通过教育评价才能获得实际意义，为改进教育的过程提供有价值的信息，否则就是简单、枯燥的数值而已。

本书中的测评，即基于证据的推断过程。[①] 测评是用来观察学生行为并产生相关数据或信息的过程，而这些数据或信息有助于对学生所知和所能形成合理推断。基于证据的推断过程，也就是收集事实及证据来支持所要做的各种推断。测评在测量和评价内涵上互有重叠。在实际运用中，往往与测量、测验、评价互为混用。[②]

教育测评表示通过专门编制或已有的工具获取学生的反应，并以此对学生的知识技能等掌握情况进行推断。在教育领域的大规模测评，也称作大规模教育测评，或大规模教育考试、大规模教育比较等，在英文名称上，常常存在“Large-Scale Educational Assessment”“Large-Scale Educational Test”“Large-Scale Survey”“Large-Scale Comparative Study”等表述。虽然“assessment”“test”“survey”“study”在使用上存在一些差异，但都指的是“对大范围（某个国家甚至几十个国家）的学生进行抽样，同时对考查内容也进行抽样的极其复杂的考试设计”。[③] 在中国，大规模测评往往是指由国家或地区的专门考试机构或行政单位（如考试院、教研室等）统一组织，在省市等级别进行的，以甄别、检测、评价等为目的的考试。大规模测评项目，简单来说就是对大量学生或者教师进行测评的项目。[④] NRC 指出，虽然大规模测评（large-scale assessments）的表达有争议，并且容易误用，但是它们在获取学生特别有价值的信息上是一个很重要的途径。大规模测评，即那些设计用来获取大量学生证据的测评，是在美国获取问责证据的基本手段。[⑤] 显然这个大规模，针对的是测评对象的数量，但是具体要达到多少并没有明确指标或数字。

目前，大规模教育测评是由专业考试机构统一设计组织，在较大范围内进行实施的，它对公平性、信度、效度等测量学特征有更为专业的要求。[⑥] 大规模测评所用工具如试卷与评分标准都是统一的，测评过程实施要求是一致的，有严密的防止舞弊的措施，有各种控制误差的技术处理，因此测评的结果是可信、可比的，特别在中国进行人才选拔时有很大的参考价值。

在教育领域，国际大规模测评，往往指的是跨多个国家（或地区）、文化的大规模教育测评。

---

① 参见：

Mislevy, R.J. . Evidence and inference in educational assessment[J]. Psychometrika, 1994, 59(4): 439—483.

Mislevy, R.J. . Test theory reconceived. Journal of Educational Measurement. 1996, 33(4): 379—416.

② Popham, W. J.. Modern educational measurement: Practical guidelines for educational reform[M]. Needham, MA: Allyn and Bacon, 2000.

③ 王蕾. 大规模考试和学业质量评价[M]. 北京：高等教育出版社，2013：38.

④ 柯政，赵小雅. 学业测量与评价的前沿和趋势[N]. 中国教育报，2013－01－02010.

⑤ Committee on Assessment in Support of Instruction and Learning, Committee on Science Education K-12, National Research Council. Assessment in Support of Instruction and Learning: Bridging the Gap between Large-Scale and Classroom Assessment—Workshop Report[EB/OL]. http://www.nap.edu/catalog/ 10802. html. 2003.11. 2015－12－6.

⑥ 丁朝蓬. 新课程改革以来学生评价改革的回顾与思考. 见：杨向东，崔允漷. 课堂评价——促进学生的学习和发展[C]. 上海：华东师范大学出版社，2012：159.

NRC指出，对于国际大规模测评，作为较大范围内跨国家的研究而言，有一个目的，就是要让我们理解在世界各地的教育中各种各样的设置意味着什么。[①] 国际大规模数学测评是指在数学教育领域内进行的国际大规模测评。

数学测评，不可视作由“数学”和“测评”两部分单独“拼成”的，必须在评价过程中考虑数学自身的特点及教学方法的实际应用。如考虑到数学的演绎证明或推理、数学在情境抽象概括中的使用和体现、数学动态变化，以及相关工具如计算机(器)等在操作使用过程中对于数学概念模型建立的影响等。数学特点和测评的有效整合，将学生作为个体来了解其学习状况，提供信息帮助教师指导、诊断学生，促进学习，直到构思或形成地区乃至国家的策略或政策，对改进地区或国家的数学教育具有重大意义，并产生深远的影响。

就数学测评而言，在具有跨多个国家(或地区)、文化的国际大规模测评中，影响最为广泛的两个分别是由OECD负责开发的PISA测评，以及由国际教育成就评价协会(International Association for the Evaluation of Educational Achievement，简称IEA)开发的TIMSS测评。这两大测评都是在国际数学教育比较研究的大背景下产生和发展起来的。

二、核心知识和核心能力

在国际大规模数学测评中，无法绕开两个方面，就是测评的内容领域及表现期望(或过程)。基于数学测评跨国的特性，测评的数学内容及表现期望要想被参加测试的国家接受或在国家之间达成共识，就涉及数学测评在核心知识和核心能力领域中的认识、确定、分类要得到公认。这些是测评目标分析框架中必须面对且非常重要的方面，可以说是整个测评设计的基础。

什么是核心知识？教育专家希尔斯(E.D.Hirsch)曾经站在课程选择的角度提出，核心知识是“共享＋稳固＋序列和具体”的知识。[②] 其他学者，如斯皮克(E.S.Spelke)[③]、孙宇浩[④]等人认为，核心知识系统出现于人类个体发展和种系发展的早期，在人类复杂认知能力的发生发展中起着建构模块的作用。这两个概念是从人类知识总体的角度上来理解核心知识的。教育专家陆建身等人认为，核心知识是“学生应该掌握的所有东西”，是“集中注意的良好教育之基础”。[⑤] 教育专家龙宝新指出，“核心知识是每个教学活动单元中必须要让学生掌握、理解、探明的主要知识技能，是一个学期教学、一个单元教学、一节课教学的主体内容与知识主干，是整个教学活动链条中的关键链环，是联系全部教学活动的主心骨，是教学活动之魂的栖息地。

---

① National Research Council. A collaborative agenda for improving international comparative studies in education[M]. Board on International Comparative Studies in Education, D.M.Gilford, ED. Commisson on Behavioral and Social Sciences and Education. Washington, DC: National Academy Press, 1993: 22.

② 赵中建.美国核心知识课程的理论和实践(上)[J].外国教育资料，1996(05)：29.

③ Spelke E. S.. Core knowledge.[J]. American Psychologist, 2001, 55(11): 1233—1243.

④ 孙宇浩，傅小兰. 核心知识系统及其对相关研究的启示[J]. 心理科学进展，2003(01)：12—21.

⑤ 陆建身.美国核心知识课程与生物学课程改革[J].生物学教学，2002(02)：9.