DOI:10.13718/j.cnki.jsjy.2015.02.002

国际教育评估与测试的研究进展*

——基于 ICME-12 的文献

周 九 诗[1],王 光 明[2]

(1. 华东师范大学 数学系,上海 200241;2. 天津师范大学 教师教育学院,天津 300387)

摘　要:评估与测试是在韩国首尔召开的 ICME-12 上的重要议题之一。基于 ICME-12,可以看到关于评估与测试的研究进展:将评估学生学业成就作为教学的有机组成部分;评估学生学业成就的方式多样;重视运用数学模型(QI 模型、PD 模型、CDMs 模型和概念评估模型)对测试题目质量进行评估。我国评估与测试的改进注意事项包括:鼓励教师和学生共同设计测试题、减少对学生解决问题时间的限制和加强评估与测试中对网络系统的使用。

关键词:评估;测试;定性评估;网络测试;自主评价;评估模型

中图分类号:G420;G633.6　　**文献标识码**:A　　**文章编号**:2095-8129(2015)02-0007-07

在我国,评估与测试对教学实践拥有强大的导向作用,关注国际数学教育界评估与测试研究的进展,有助于促进我国数学评估与测试的改革。在韩国首尔召开的 ICME-12(第 12 届国际数学教育大会)上,将评估与测试作为一个重要专题进行了讨论,来自 19 个国家和地区的 71 名学者参与了此讨论,展现当今国际视野下,各国数学教育工作者在评估与测试方面的研究进展与成果。本文基于 ICME-12 中关于评估与测试的文献,阐述国际数学教育界关于评估与测试的研究进展,进而剖析我国数学教育测试与评估的不足,并提出改进意见。

一、将评估学生学业成就作为教学的有机组成部分

评估学生的学业成就是教与学不可分割的一部分[1],在整个教学过程中起着积极作用。评估学生学业成就的作用具体表现在以下几个方面:

(一)诊断功能

对学生的学业成就进行评估是对教师教学途径和教学方法的一种考核方式,合理运用评估结果能够使“教”与“学”不断臻于完善。在 ICME-12 上,与会代表达成共识:评估与测试的根本目的是改善教师的教学方法和学生的学习方式。学生学业成就的评估是对教学过程和结果的一种分析,借此了解教与学中存在的问题[2],对教与学进行诊断。但诊断的目的并不仅在于对学生的学习

* **收稿日期**:2014-12-26

作者简介:周九诗,华东师范大学数学系博士研究生。
王光明,天津师范大学教师教育学院院长,教授,博士生导师。

基金项目:教育部人文社科一般项目“高中生高效学习的心理特征研究”(13YJA190012),项目负责人:王光明;天津市高校“中青年骨干创新人才培养计划”(“十一五”期间),项目负责人:王光明。

成就进行筛选和区分等级，而是要判断教师的教学是否有效、教学进度是否合理、教材选用是否恰当、学生对知识的掌握是否牢固等。

(二)促进功能

对学生的学业成就进行评估不仅能肯定和强化有效的教学方法和先进的教学理念，还能对学生的知识学习起到促进作用[2]。一方面，学生可以从评估中肯定自己的学习成就，进而激发内在学习动机；另一方面，学生能够根据评估结果，发现自身存在的知识薄弱点，从而在今后的学习中，对掌握不牢固的知识点进行强化和巩固，或对不恰当的学习方法加以改进和完善。

(三)调整功能

学生学业成就的评估结果所提供的反馈信息，可以作为提高学生学习效率和改进教师教学方法的依据[3]，帮助教师和学生改进数学教学中“教”和“学”的方式[4]，师生对评估过程、结果进行共同探讨，体会因评估所带来的教学方法和学习方式的改变。评估与测试应该理解为一种过程、一项持续的活动，而不是一种最终结果。以学生为主体、促进学生学习，相应调整教学内容和教学方式是评估的价值取向。

二、评估学生学业成就的方式多样

对学生学业成就进行的评价不能只注重测试结果，关注教学目标的达成情况，只了解学生知道或学会了什么，而不了解学生的思考过程。优良的学生学业评价方式应该能够从多个角度、全方面地评估学生的表现，高质量的评估不仅要注重评估结果，更要注重评估过程[5]。

(一)重视定性评估

定性评估采用更人性化的评价方式，全面分析学生出错的原因，进而更加合理地评估学生学业成就。能够采取定性方式评估的题目大多属于主观性开放题，这也使得开放性题目成为很多国家测试命题的一种趋势。一方面，开放性问题大多没有最终唯一的标准答案，能激发学生的想象力，培养学生的发散思维和对知识灵活运用的能力；另一方面，主观性开放题较客观题来讲，更能反映学生的思考过程，使教师有效掌握学生的解题思路，全面评估学生的学业成就。

新加坡国家教育部提出的 SMAPP(The Singapore Mathematics Assessment and Pedagogy Project)项目，即新加坡数学评估和教学法项目，共设置了 8 个情景化问题，其中一些问题已被用于新加坡中学教科书的编写中。8 个情景化问题都属于开放性问题，包含多个小问，对每一题的评估都采用定性与定量相结合的方法，从 5 个方面评价学生的学业成就。这 5 个方面包括：理解问题和提取信息、计算、推理证明、表征和用数学方法解决实际问题的能力[6]。并不是所有问题都能全面地考核这 5 个方面，每个问题具体涉及到的考核部分都会有明确标注。同时，研究也表明，如果评估过程能够加入学生访谈，效果会更好。

Kym Fry 等提倡用调查性数学问题来评估学生学业的成就。调查性数学问题中的题目大部分为学生感兴趣的话题，多数为开放性问题和结构不良的问题，旨在培养学生的数学素养。调查性数学问题的解题方法不局限于课堂上教师的讲解，还可能会用到课外知识。调查性数学问题以学生为中心，采取小组合作形式，教师在其中只起辅助作用，教师和学生一起商讨学习内容和学习方式，并不过分强调对概念理解的练习强度和记忆能力。调查性数学问题分为 4 个阶段：理解问题—制定解题计划—完善计划—得出结论并解释结果。教师会根据学生每一阶段的表现给出相应的定性评估[7-8]。譬如，“我们的教室有多整洁?”就是一个调查性数学问题，学生要收集相关数据，确定

调查研究的方法，根据对数据的解释分析得出结论，判断方案的数学合理性。这一过程一般需要学生小组合作 6～8 小时完成，教师会给学生 6 天左右的时间去解决问题。

Bell 等认为描述性测试是评估学生学业成就的理想工具之一。区别于多选、判断正误或简答形式的传统测试，描述性测试的重点并不是考查学生的考试技巧和记忆力，而是更多地分析学生的解题过程和如何利用评估结果提高教学质量。Gearhart 和 Saxe 认为，教师“知道学生所知道的”是有效教学最重要的一部分。目前韩国中学的期中、期末考试大多采用描述性测试，TIMSS 2007 中有 45.6%的问题都是描述性测试问题。描述性测试采用开放的问答方式，通过学生写出他们是如何理解问题和解决问题的方式来评估学生的思考过程，评价体系分为 A、B、C、D 4 个等级，分别为优秀、良好、合格和较差，每个等级标准都有相应的描述性量规，相对定量评估更加全面和灵活[9]。

Martha J.koch 等从社会文化角度考虑数学的教与学，他们认为教师的教、学生的学和对学生学业成就的评估是融为一体的。研究表明：采用统一、标准的量规化评价方式会扼杀学生的想象力，限制学生数学能力的发展。教师不是仅完成已有课程标准要求的“工匠”，应鼓励教师根据题目特点建构满足学生需要、符合题目测试目的的评价量规，借以多角度、全面、有针对性地评估学生学业成就。教师还可以通过和同事讨论、和学生谈话的方式改进评价量规[10]。

(二)加强网络测试

在科技迅猛发展的今天，现代化信息技术与数学学科的整合已成为一种不可抗拒的趋势。网络性测试相对于传统的测试方式更加方便快捷，且不受时间和空间的限制。一方面，网络测试所具有的便捷资源共享途径，使其测试资源更为丰富多样；另一方面，测试系统对学生解答反馈的及时性，有助于提高教学效率。ICME-12 的报告中共介绍了两种数学网络评估系统，分别是 smart(specific mathematics assessment that reveals thinking)测试和 STACK 测试。

smart 测试系统的测试内容涵盖澳大利亚初中数学课程的大部分知识点，在澳大利亚的中学里具有较高的使用频率。这个系统能够在学生回答完毕后，及时将学生出错的原因和学生对问题的理解程度反馈给教师，并提出有针对性的教学建议。通过 smart 测试系统，教师可以直接洞察学生的思考过程。教师使用 smart 系统的流程如下：教师首先选取测试主题，之后启用有关此主题的 smart 测试题目并由学生作答，然后通过 smart 测试给出的解题反馈，了解每个学生的解题方式，随后教师采用 smart 系统提供的新题目来继续测试学生，新的测试题目往往在原先题目的基础上作了微小变动，目的是为了更准确地判断学生对题目的掌握程度，及时发现学生的知识漏洞[11]。

STACK 测试和 smart 测试类似，这两种测试系统都可以根据学生的不同回答及时给出有针对性的反馈。和 smart 测试相比，STACK 测试可以自动生成随机问题，问题库更加丰富多样，反馈报告也更为详细。由于计算机识别能力的局限性，网络测试往往很少出填空题或解答题，譬如 smart 的测试题目就以选择题(多选或单选)为主，但 STACK 测试的识别系统十分强大，可以准确判断学生给出的答案在数学本质上是否正确，如一道填空题目的答案是 1，学生如果填 7°，STACK 测试也会认为答案正确。STACK 测试可以在一定程度上缓解网络测试存在的机械性与生硬性的问题。目前，STACK 已被译为多种语言(芬兰语、葡萄牙语、德语、荷兰语和日语)的版本，在全世界范围内使用[12]。

(三)强调自主评价

Theodora Teong Ying Xi 等提倡学生使用“自我评价包”促进自我监督性学习，使学生积极参

与自我表现的评价。在此过程中学生能够更为准确地了解自身对知识、技能的掌握程度，改进所存在的不足。不仅如此，教师还可以通过学生的自我评价，了解所教内容与学生实际学习情况的差距，帮助学生了解“要解决什么问题”、“如何解决这个问题”和“解题的下一步要做什么”。“自我评价包”包括 3 部分：知识点总结、自我评价打分表、自我小结（如自我评价和教师评价存在差异的原因和如何提高自我打分表的成绩等）。为了进一步满足不同学生的需要，在学生使用一段时间“自我评价包”后，其具体评价内容可由教师根据学生自身特点进行改进。Cheng Lu Pien 的调查进一步证实，使用“自我评价包”的学生比不使用“自我评价包”的学生在数学能力方面提高得更快[13-14]。

三、重视运用数学模型来评估测试题目的质量

如何设计优良的测试题目是数学教育者不断探索的一个问题，高质量的题目有助于提高评估效果。对于如何判断评估题目质量的问题，ICME-12 的报告中介绍了几种设计题目和评估题目的数学模型，值得参考借鉴。

（一）QI 模型（质量指数模型）

Belinda Huntley 介绍了一种评估题目质量优劣的模型——QI 模型[15]。QI 模型有 4 个参数：题目区分度、学生对题目的自信度、学生答题的预计表现和题目难度。其中，前 3 个参数的数值可以在三维坐标 x，y，z 轴上围出一个区域，这个区域面积用 QI 表示，若 $QI<0.282$，则为高质量题目；反之，则为低质量题目。题目难度则反映在这个所围区域颜色的深浅上，颜色越深说明题目越难。譬如题目的区分度为 0.295，学生对此题的自信度为 0.385，专家对学生答此题的预计表现为 0.236，则根据这 3 个数据依次在三维坐标轴标出 3 点(0.295，0，0)，(0，0.385，0)，(0，0，0.236)，由这 3 点围成的三角形区域面积为 0.119，即这道题目的 QI 值为 0.119，由于 $0.119<0.282$，说明此题是一道“好”题。

（二）PD 模型（技能发展模型）

David C.Webb 认为，PD 模型可以用于提高中学数学的课堂评估，PD 模型比较适合于评估比例推理和代数问题[16]。教师的自身知识水平直接影响测试题目的质量及测试效果，PD 模型能够引导教师设计优质题目。教师使用 PD 模型前要改进之前对传统评估概念的理解和观念，PD 模型为教师提供重新理解教学目标的契机，帮助教师重新评估考核对教学目标完成度所达到的程度，为教师选择、设计测试内容提供一双“设计者的眼睛”。PD 模型的流程如下：首先由教师设计基本测试内容，测试之后教师会分析并解释学生的表现，然后和其他教师共同讨论如何适当改进测试内容和教学计划，最后制定出一份小规模的形成性测试。PD 模型不是线性结构，而是一个循环，即教师制定出一份小规模的形成性测试后再重新投入测试，并分析学生对新测试内容的表现，从而不断完善测试内容。

（三）CDMs 模型（Cognitive Diagnosis Models）（认知诊断模型）

Hartono Tjoe 等建议使用 CDMs 模型设计比例推理问题[17]。CDMs 模型是一种心理测量模型，从 6 个方面考查比例推理问题：比例技巧、分数比大小、建构比例、理解变量之间的比例关系、判断两个变量是否存在比例关系、用算法解比例问题。基于 CDMs 模型设计的题目，可以同时考查学生对多个知识点的掌握情况，评估出学生知识薄弱的方面，并针对这一方面给予及时反馈和改进建议。同时，CDMs 模型还可以根据学生的测试表现，推断出学生完成测试还需掌握的知识点。譬

如：Emma 和 Irene 每人有 5 个苹果，Penny 的苹果比 Emma 的苹果数少 1；Penny 和 Irene 每人有 7 个樱桃，Emma 的樱桃比 Penny 的樱桃数少 1。将每个人手中的苹果数比樱桃数的结果由小到大排序，下列选项正确的是？ A.Emma，Irene，Penny；B.Irene，Emma，Penny；C. Emma，Penny，Irene；D. Penny Irene，Emma。此题目同时从两个方面考查了学生掌握比例推理问题的情况，一是建构比例的能力（如何表示每个人手中的苹果数和樱桃数的比），二是比较多个分数大小的能力。

（四）概念评估模型（Conceptual Understanding Assessment）

Mehmet Turegun 介绍了一种用于评估学生对统计概念问题掌握情况的模型[18]。Cobb 认为，传统的统计概念测试将重点放在计算能力的考核上，并没有评估学生的高层次认知和概念理解情况，概念评估模型可对这一现象进行有效改进。概念评估模型的理论基础来自皮亚杰的认知发展理论和建构主义，基于 SOLO 分类评价法，将学生的概念理解水平分为 5 类：前结构层次、单点结构层次、多点结构层次、关联结构层次和拓展抽象结构层次，详细描述每个层次的划分标准，教师可以参照此标准就学生对该概念掌握的情况作出判断。具体划分标准如下：前结构层次（收集的信息和问题无关或对问题的理解有误）；单点结构层次（依赖对概念定义的讨论和描述，无法给出不同解题方法间的相关联系）；多点结构层次（能够给出概念定义，对于概念的外延和内涵能给出模糊性的界定，能够比较不同解题方法的异同和联系）；关联结构层次（能够比较不同解题方法的异同和联系，能够解释中值、平均数和异常值）；拓展抽象结构（在解决复杂问题时，知道哪种解题方法最合适）。

四、评　述

从 ICME-12 中评估与测试的相关报告看出：当今国际数学教育界把评估作为提高教学质量的重要手段；合理有效的评估能够对教师的自我完善和学生的全面发展起到积极作用；评估学生学业成就的目的是探寻学生薄弱的知识点，调整教师的教学方法；对于测试而言，不是任何数学测试题都是质量优良的，需要对数学测试题的质量进行评估。透过 ICME-12 中关于评估与测试的研究成果，以国际数学教育界关于评估与测试的理念与实践做法作参照，发现我国数学教育实践中的一些不足。

首先，尽管我国数学课程标准明确指出注重对数学学习过程的评价，但是，我国教学的实践并没有充分重视过程评价。之所以会出现这种情况，应试教育固然是主要原因，但可操作性的过程性评价研究成果缺乏，对实践指导不力也是重要原因。

其次，我国基础教育实践评估与测试评价中（尤其是中学）重测试，但缺乏评价测试题好坏的数学模型的运用。我国试题编制人员以及教师应运用类似 QI 模型，设计出质量优良的试题，使对测试题好坏的评价不再见仁见智，而是由公认的评价模型去作仲裁。

再次，过于重视统一、标准化的测试。考试管理部门以及教师要认识到有些数学能力以及数学素质通过测试是测量不出来的。Martha J. koch 等甚至认为采用统一、标准的测试会扼杀学生的想象力。IB 课程评价引导学生阅读数学（文献）、理解数学、建构数学、应用数学、分享数学的做法也值得我国数学教育工作者借鉴。

最后，缺乏网络测试系统。网络测试的便捷性和多样性使得越来越多的网络测试程序问世，这也是评估测试的一种新兴模式，我国亟待开发适合我国国情的测试系统。

五、建　议

针对我国测试与评估方面存在的不足，我国数学教育测试与评估亟待改革，改革中应该注意的

事项包括：设计评估与测试的题目时应更多考虑学生的自身特点，由师生一起参与题目的选择和编写；对题目的解决时间也可以适当放宽，争取更加全面地考查学生的学习成就并借此提升学生的数学素养；充分利用信息技术的便捷性和广泛性，加强评估与测试中对网络系统的使用。

(一)鼓励教师和学生共同设计测试题

测试题目的范围不应局限于课本，测试形式也应灵活多样。国际上越来越倾向于使用开放性问题和结构不良性问题来评估学生的学习成就。教师可以征求学生意见，选择学生感兴趣且有价值的专题来安排学生学习，激发学生学习的主观能动性，使学生在研究问题时投入更多的热情和精力，客观正确地评价学生的学习水平。

(二)适当放宽对学生解决问题时间的限制

很多国际学者都认为，一道数学题学生会做就可以了，为什么强调必须在一定时间内(往往很短)完成？好的数学题往往是无法在短时间内完成的。譬如，解一道调查性数学问题一般需要学生合作 6～8 小时，教师会给学生 6 天左右的时间去完成问题。目前我国教科书中也涉及到一些开放性问题，但是对这些问题的掌握程度的评估还主要依赖限时的纸笔测试，而几乎很少合理利用这些开放性问题来训练学生的数学思维和提高学生的数学素养，因此，建议教师合理安排学生小组合作，给学生更宽裕的时间以完成开放性问题。

(三)逐步推广使用网络测试系统

在科技飞速发展的大背景下，我国课程标准也十分强调教学中对信息技术的使用。网络测试评估系统可以提高学生自我评估的效率，有助于学生在课堂之外的地方学习。网络测试系统能够为学生提供及时的反馈并给出有针对性的教学建议，虽然人性化稍有欠缺，但质量优良的网络测试还是能够给出详细具体的反馈报告，帮助学生认识自身不足，找出自身知识的薄弱点。我国人口基数较大，各地教学质量不一，在网络较为普遍的今天，开发和使用网络测试也能够在一定程度上缓解一些地区教育普及程度较低和教学质量较差的状况。

参考文献：

[1] Liora H，Miriam A.Assessment and Evaluation-the Link Between The Knowledge and Behaviour of Novice Mathematics Teachers [C].Seoul：International Commission on Mathematical Instruction，ICME-12，2012：6563.

[2] Heidi K，Laura K，Jari L. Assessment in Finnish Mathematics Education：Various Ways，Various Needs[C]. Seoul：International Commission on Mathematical Instruction，ICME-12，2012：6662.

[3] King M，Leung.Using Large-Scale Assessment Data to Enhance the Teaching of Mathematic[C].Seoul：International Commission on Mathematical Instruction，ICME-12，2012：6671.

[4] Patricia D H，Denisse R.The Extent to Which Primary Assessment the U.S. Engage Students in Representation[C].Seoul：International Commission on Mathematical Instruction，ICME-12，2012：6581.

[5] Kim R Y，Kim K Y.The Challenges and Issues Regarding Extended Constructed-response Questions：Korean Teachers'Perspective [C].Seoul：International Commission on Mathematical Instruction，ICME-12，2012：6631.

[6] Cheang W K，Teo K M，Zhao D S.Assessing Mathematical Competencies Using Disciplinary Tasks[C].Seoul：International Commission on Mathematical Instruction，ICME-12，2012：6504-6513.

[7] Kym F，Katie M.Assessing for learning in inquiry mathematics[C].Seoul：International Commission on Mathematical Instruction，ICME-12，2012：6534-6543.

[8] 宋广文，李晓芹，朱振菁．小学儿童数字线估计的心理表征模式[J]．数学教育学报，2013，22(5)：52-56.

[9] Kim M K.Development and Implication of Descriptive Assessment in Elementary Mathematics Classroom in Korea[C].Seoul：International Commission on Mathematical Instruction，ICME-12，2012：6621-6630.

[10] Martha J K, Christine S. Teachers Working Collaboratively to Further Develop Their Assessment Practices in Mathematics: Turning Rubrics Into Non-rubrics[C]. Seoul: International Commission on Mathematical Instruction, ICME-12, 2012: 6641-6650.

[11] Vicki S, Kaye S. Teachers' Views of Using an On-line, Formative Assessment System for Mathematics[C]. Seoul: International Commission on Mathematical Instruction, ICME-12, 2012: 6721-6730.

[12] Christopher S. Computer Aided Assessment of Mathematics Using Stack[C]. Seoul: International Commission on Mathematical Instruction, ICME-12, 2012: 235-250.

[13] Theodora T Y X, Cheng L P. Developing Self-regulated Learners Using Self-assessment in the Primary Mathematics Classroom[C]. Seoul: International Commission on Mathematical Instruction, ICME-12, 2012: 6514-6522.

[14] 何声清,巩子坤. 11－14 岁学生关于可能性比较的认知发展研究[J]. 数学教育学报, 2013, 22(5): 57-61.

[15] Belinda H. What is A "Good" Mathematics Test Item? [C]. Seoul: International Commission on Mathematical Instruction, ICME-12, 2012: 6591-6600.

[16] David C W. Teacher Change in Classroom Assessment: the Role of Teacher Content Knowledge in the Design and USE of Productive Classroom Assessment[C]. Seoul: International Commission on Mathematical Instruction, ICME-12, 2012: 6773-6782.

[17] Hartono T, Jimmy de la T. Proportional Reasoning Problems: Current State and a Possible Future Direction[C]. Seoul: International Commission on Mathematical Instruction, ICME-12, 2012: 6741-6750.

[18] Mehmet T. Developing an Alternative Mode of Assessment for Conceptual Understanding[C]. Seoul: International Commission on Mathematical Instruction, ICME-12, 2012: 6751-6763.

A Review of Researches on International Education Evaluation and Testing in ICME-12

ZHOU Jiu-shi[1], WANG Guang-ming[2]

(1. *Department of Mathematics, East China Normal University, Shanghai* 200241, *China*;
2. *College of Teacher Education, Tianjin Normal University, Tianjin* 300387, *China*)

Abstract: Evaluation and testing is one of the most essential topics at the ICME-12 Conference held in Seoul, South Korea. Based on ICME-12, we may find the progress made in the research of evaluation and testing, including subsuming the evaluation of students' academic achievement in teaching, diversifying evaluation approaches to students' academic achievement, and valuing the application of mathematical model(QI model, PD model, CDM model and conceptual assessment model) to quality assessment of test items. The paper gives some suggestions: encouraging teachers and students to design test together, giving students more time to problem-solving and enhancing the application of internet in evaluation and testing.

Key words: evaluation; test; qualitative evaluation; online test; independent-evaluation; evaluation model

责任编辑　唐益明