雷新勇 著

基于标准的教育考试

—— 命题、标准设置和学业评价

本书以大规模基于标准的教育考试为主要研究对象，讨论国际基于标准的教育考试的发展现状和趋势、考试效度和效度检验的历史、最新理论成果以及基于标准的教育考试效度检验框架、考试命题、标准设置、信度研究和考试结果成绩报告。

上海科学技术出版社

基于标准的教育考试

——命题、标准设置和学业评价

雷新勇 著

上海科学技术出版社

内 容 提 要

本书是讨论大规模基于标准的教育考试命题、学业标准设置和学业评价问题的一本实用性理论著作。

全书共十五章。

第一章主要介绍基于标准的教育考试与常模参照考试的区别，基于标准的教育考试产生的国际背景及其特征。

第二章介绍基于标准的教育考试的效度检验的最新理念、方法和要求，以及这些理念、方法和要求对基于标准的教育考试的设计、开发的启示。

第三章介绍大规模基于标准的教育考试需要考虑的基本问题。

第四章介绍大规模基于标准的教育考试中学业标准的开发。

第五章到第八章主要介绍基于标准的教育考试的试题的命制。

第九章介绍主观题评分标准的开发。

第十章讨论大规模基于标准的教育考试的内容选择和组卷要求。

第十一章到第十三章介绍标准设置。

第十四章介绍大规模基于标准的教育考试的信度。

第十五章介绍大规模基于标准的教育考试的考试分数报告。

图书在版编目(CIP)数据

基于标准的教育考试：命题、标准设置和学业评价/雷新勇著. —上海：上海科学技术出版社，2011.4

ISBN 978-7-5478-0697-5

Ⅰ.①基… Ⅱ.①雷… Ⅲ.①考试—标准—研究 Ⅳ.①G424.74

中国版本图书馆 CIP 数据核字(2011)第 031350 号

责任编辑 邵海秀

上海世纪出版股份有限公司
上 海 科 学 技 术 出 版 社 出版、发行
(上海钦州南路 71 号 邮政编码 200235)

新华书店上海发行所经销
苏州望电印刷有限公司印刷
开本 787×1092 1/16 印张 21
字数：383 000
2011 年 4 月第 1 版 2011 年 4 月第 1 次印刷
ISBN 978-7-5478-0697-5/G·130
定价：42.00 元

序　言

在我们即将走进“十二五”，开始按照《国家中长期教育改革和发展规划纲要》的部署进行新的教育改革和发展之际，我非常欣喜地看到雷新勇同志的新作《基于标准的教育考试——命题、标准设置和学业评价》基本完稿。

2001年，教育部颁布的《基础教育课程改革纲要（试行）》明确指出：“国家课程标准是教材编写、教学、评估和考试命题的依据，是国家管理和评价课程的基础。”这指出了21世纪以来我国开展的基础教育领域的新课程改革，本质上是基于标准的教育改革。2006年，随着基础教育领域的新课程改革的深入，考试评价的改革提到了教育改革的议事日程上。2007年，《教育部关于普通高中新课程省份深化高校招生考试改革的指导意见》指出：“各地要加快建设在国家指导下由各省份组织实施的普通高中学业水平考试和学生综合素质评价制度，逐步发挥其对普通高中教育教学质量进行管理和监控，对高中学生学业水平和综合素质进行全面、客观评价以及为高校招生选拔提供参考依据的作用。”基础教育的改革者看到了考试评价改革的必要性和迫切性。在此前后，一些省市开始考虑设置高中学业水平考试，一些省市也将已经实施多年的高中会考改名为高中学业水平考试。在此之前，一些省市也对初中毕业考试和高中阶段招生考试实施改革，实施了初中学业水平考试。上海市教育考试院2005年开始在教育行政部门的指导下，具体研究设置高中学业水平考试的诸方面问题。

在这样的大背景下，作为我院命题工作的负责人，雷新勇同志于2005年就开始了基于标准的教育考试的专门研究。我们经常在一起讨论有关学业水平考试的诸多问题：

为什么要设置学业水平考试？

学业水平考试与会考有什么区别？

学业水平考试应该以课程标准为依据，究竟以课标的什么内容为依据？如何根据课程的内容标准组织命题？

如何确定考生的学业水平？

如何根据学业水平考试结果对考生的学业进行评价？如何根据学业水平考试结果对群体进行评价，如对学校的教学水平、甚至区县的教学水平进行评价？

如何根据学业水平考试结果诊断考生的学习弱项？如何根据学业水平考试结果诊断班级、学校和区县教学存在的问题？如何根据学业水平考试结果对课程进行评价？

这些，时而在正式场合研讨，时而是茶后闲谈交流的问题，激发了雷新勇同志学术研究的热情，如今，点滴思考已汇成系统的理论和实践成果。

雷新勇同志不满足于仅仅进行基于标准的教育考试的理论研究，其在理论研究的同时，始终不断地将理论研究结果应用于考试实践。在上海市初中学业水平考试命题中，他不断地向命题教师介绍基于标准的教育考试命题的新要求，同时也向已经实施的上海市高中学业水平考试学科命题教师介绍基于标准的教育考试命题的新要求。在部分初中学业水平考试学科中组织实施标准设置，探索在我国基于标准的教育考试中实施标准设置的可行性，并发现其中需要注意的问题。

2008年，雷新勇同志又担任了全国高等教育自学考试上海命题中心主任，高等教育自学考试，本质上也是基于标准的考试。这使得雷新勇同志有机会以更加宽广的视角考察我国基于标准的考试中的问题，思考解决之道。

长期的理论研究和教育考试实践催生了这部新作《基于标准的教育考试——命题、标准设置和学业评价》，它比较系统地回答了初中学业水平考试、高中学业水平考试、高等教育自学考试的考试设计、命题、标准设置、学业评价的相关问题，从一个侧面反映了上海市教育考试院在基于标准的教育考试方面的研究水平。

本书第一章介绍了基于标准的教育考试产生的国际背景及其特征，这之中令我印象最深的是“基于标准的考试成为推动教育改革的有效工具”。在基于标准的新课程改革中，我们确定了课程标准对课程、教学、考试评价的指导作用，然而在实际课程改革的推进中，有些人不承认考试对教育改革的推动作用，试图从各个方面抑制考试功能的发挥。从国际基于标准的教育改革经验中，我们看到基于标准的考试已经或正在成为教育决策者推动教育改革的工具。在实践中我们看到的事实是，大规模教育考试对教育教学的影响是无容置疑的。我们与其不接受，倒不如因势利导，充分发挥基于标准的教育考试的作用，推动教育改革深化。问题是如何达到这个目的？本书第一章概括了国际基于标准的教育改革和考试评价发展的经验，告诉我们保持基

于标准的教育考试与课程内容标准调整一致，广泛采用表现性评价，向考生、家长和学校等报告学生的学业水平及学习中存在的问题，可以有效地推进课程改革，改善学校的教学，改善学生的学习。本书就是以基于标准的考试与课程内容标准的一致性为主线，展示对基于标准的教育考试理论和实践的探索。

《国家中长期教育改革和发展规划纲要》指出："深化考试内容和形式改革，着重考查综合素质和能力。"然而，对于考试内容究竟是什么，许多人并不清楚。大家看了PISA(国际学生评估项目的缩写)的试题，认为PISA考试好，我们的考试内容改革要向PISA学习。其实PISA考试的本质，强调的是基本的知识内容及应用知识解决问题的认知技能，这是PISA考试内容的核心。本书第五章到第十章从教育测量的角度告诉我们，考试内容就是考生学习的知识内容及其认知要求，告诉我们应该如何以课程内容标准为依据，命制试题、选择考试内容、组成试卷，也告诉我们怎样命制试题，来考查考生的认知能力。

主观题评分标准的开发是本书的一大特色。国际基于标准的教育改革中一大重要的改进就是考试评价中广泛采用表现性评价，即主观试题，主观题开发中一个重要方面就是评分标准的开发。当我们很多人还在热衷于主观题的采点评分的时候，国际基于标准的教育考试早已进入了主要特质评分法的时代，PISA、TIMMS等考试的评分采用的就是主要特质评分法。其实这也是考试内容改革的重要方面。近年来，上海市教育考试院在高考主观题中开始采用主要特质评分标准，这也是考试内容改革的一个重要方面。本书第十章告诉我们主要特质评分标准的意义和作用，如何开发主要特质评分标准。

基于标准的考试与以往的会考的一个重要的区别是前者根据考生达到课程内容标准规定的学习要求的程度，确定考生的学业水平；而后者则根据比例划分等第。前者是以课程标准为依据的，后者是人比人的。长期以来，很多人并不知道如何将定量的考试分数量表映射到定性的考生学业表现量表上，来确定考生的学业水平，以为按照比例划分等第是确定学业水平的唯一方法。本书第十一章到第十三章以标准设置为题，介绍了依据学业表现确定考生学业水平的理论和方法。

学业水平考试的结果最终要用于对考生的教育决策——确定考生的学业水平，诊断考生学习中存在的问题和优势，并且根据考生群体的学业水平和学习的弱势，推测学校教学中存在的问题，对群体做出评价。所有这些结果都需要以成绩报告的形式反馈给考生、考生家长、学校、乃至不同级别的教育行政机构。考试机构应该如何报告考生的学业水平、如何报告学生学习的强项和弱项，我们过去在这方面研究甚少，几乎完全忽略了成绩报告单的评价和诊断功能。本书第十五章以分数报告为题，介绍如何以成绩报告单的形式，向不同的对象报告不同的评价和诊断结果。非常值得学习和借鉴。

读罢首稿，掩卷沉思。教育招生考试机构减少行政色彩向专业化方向转变，为考生、学校和社会提供更加专业的考试服务，是未来教育考试机构改革的必由之路。教育考试机构的专业化发展，不但需要造就一支专业化的学科秘书队伍，还需要教育考试机构的领导具备一定的教育测量学素养，教育考试机构的其他从业人员掌握一定的教育测量学知识。这样，才能在教育考试的各个环节——不仅仅是命题环节——按照教育考试的客观规律和科学要求进行考试的设计、命题、施考、数据处理、学业评价和结果报告，进行考试评价。也需要教育行政机构的领导具备一定的教育测量素养，科学地领导基于标准的课程改革和考试评价。

《基于标准的教育考试——命题、标准设置和学业评价》一书为教育考试机构专业化发展，为教育考试机构工作人员提高教育测量学素养提供了一个很好的学习读本，我期待《基于标准的教育考试——命题、标准设置和学业评价》早日出版。

上海市教育考试院院长、教授

李瑞阳

2011年2月

前　言

进入21世纪以来，随着我国基础教育课程改革的推进，颁发了新的学科课程标准。教育部明确：课程标准是教材编写的依据，是指导课堂教学的指南，也是考试评价的依据。根据这个规定，近几年来，许多出版社依据新课程标准，先后编撰和出版了一系列新教材；在新课程标准的指导下，基础教育领域的课堂教学也日渐发生令人惊喜的变化；评价改革也在逐步推进。作为新课程评价改革的一个重要组成部分，一些省、市先后推出了初中学业水平考试和高中学业水平考试。2008年1月，《教育部关于普通高中新课程省份深化高校招生考试改革的指导意见》（以下简称《指导意见》）规定："各地要加快建设在国家指导下由各省份组织实施的普通高中学业水平考试和学生综合素质评价制度，逐步发挥其对普通高中教育教学质量进行管理和监控，对高中学生学业水平和综合素质进行全面、客观评价以及为高校招生选拔提供参考依据的作用。"根据这一要求，部分省、市将过去的高中会考，更名为高中学业水平考试。从一些省、市教育厅或教委下发的文件来看，所有省、市毫无例外地将学业水平考试定义为标准参照考试或基于标准的考试，标准就是新课程标准；学业水平考试的结果是评价学生学业水平的重要依据，也是评价学校教学质量的依据。这一系列的改革步骤说明，新课改本质上是基于标准的教育改革。

基于标准的教育至少包括四个要素：课程标准、教材、课堂教学、考试评价。这些要素整合成一个整体，互相配合，指导学生学习标准规定的学习内容，从而达到规定的要求。要达到这一目的，重要的技术要求就是基础教育体系中的教材、课堂教学、考试评价等要素，必须与课程标准保持一致；如果

不一致，这些要素就需要调整，以达到诸要素与课程标准一致的要求。教材、课堂教学、考试评价与课程标准之间的非常清晰的一致性关系，可以驱动课程改革向期望的方向发展。

学业水平考试受制于基于标准的教育体系的基本要求；同时，由于它属于教育测量学范畴，也应受到现代教育测量学基本原理的制约。从这两个视角来审视我国各省、市现已实施或即将实施的学业水平考试，可以发现其中存在的一系列问题，这些问题的存在也有着深刻的历史原因和现实原因。

一、我国学业水平考试存在的主要问题

1. 学业标准缺位

各省、市推出的学业水平考试，毫无例外地规定要用等第制来报告学生的学业水平，如优秀、良好、合格、不合格，或 A、B、C、D 等。然而，检查新课改的课程标准或多数省、市的学业水平考试纲要，不难发现：新课程标准没有就学生的学业水平确立标准，很多省、市的学业水平考试也没有学业标准。各个省、市对学生进行分类决策的依据都是学生的百分位分数。以学生百分位分数为依据，从本质上来说，是人比人，是以其他学生的表现为依据衡量学生的学业，这是常模参照比较。显然，以百分位分数作为对学生进行分类决策的依据是不合适的，因为它与基于标准的教育改革的本质目的不相符。殊不知，在基于标准的教育体系中，学业水平考试是评估学生在经过一定时间的学习后，所达到的学业状态或水平，这个状态或水平的评估只能以课程标准的要求为参照。

各个省、市的学业水平考试之所以采用百分位分数作为分类决策的依据，一个重要的原因就是没有学业标准。

学业标准从哪里来？笔者认为，最好是在制定课程标准的过程中，既制定内容标准，又制定学业标准（又称为表现标准）。然而令人遗憾的是，我国所有的课程标准都只有内容标准，没有学业标准，这是我国学业水平考试的致命伤。没有学业标准，如何确定学生的学业水平？这就很难保证学业水平考试能够与课程标准、学校教学保持一致。

2. 考试内容选择的思路欠清晰

新课改经常伴随着人们对教育考试的批评，其中最多的是来自教育行政机构领导对考试内容的批评。他们认为考试内容陈旧，需要改革。然而，对于考试内容改革究竟改什么？改革要解决什么问题？如何选择考试内容？尚没有足够的理性思考。

其实，考试内容的改革需要考虑多种因素。就学业水平考试而言，其中最重要的因素包括：考试目的、课程标准、根据考试结果所作出的推测、考试的效度或有效性。在此基础上，笔者认为，学业水平考试的内容选择应遵循以下四个原则。

(1) 内容领域或主题的一致性原则。学业水平考试所规定的内容领域或主题，应该与课程标准所规定的内容领域或主题保持一致。只有两者高度一致，学业水平考试对学校教学的反拨作用才能得以发挥；学校才会按照课程标准所规定的内容领域或主题组织教学活动，才会向学生报告他们在不同内容领域或主题上的学习成果，才会为他们提供有关知识内容学习情况的诊断信息。

(2) 知识深度与水平的适应性原则。试题考查的知识内容的深度或复杂度，应该与课程标准所规定的相应知识的认知要求相适应。而学业水平考试测量学生在内容领域或主题上的学习结果，也应该基于学生对该内容领域或主题专门知识的理解程度。由此可见，基于标准的考试必须对考查的内容领域或主题知识、技能的深度作出规定，否则，命制的试题可能会偏离内容标准的要求。

(3) 内容主题的知识范围对应性原则。该原则要求考试测量的内容范围与课程标准规定的内容范围对应一致，即考生正确应答试题所需要的知识范围，应与标准期望的学生学习的知识范围相同或对应。任何一个考试，由于时间和题量的限制，只能是抽样测试。从理论上来说，如果试题样本具有足够的代表性，知识范围的对应性则不成问题。然而，在我国所有教育考试的设计中均很少考虑教育测量学的基本原理，考试的题量通常偏少，试题样本的代表性不足。

(4) 内容主题或知识的平衡性原则。该原则要求考试内容规范要考虑内容主题或知识点的权重。学业考试内容规范不但要求试题知识深度水平(内容复杂性)和涉及的知识范围的规定与课程标准的内容要求可以进行对比，而且要求不同内容领域或主题，以及每一主题下知识点的权重也应该与课程标准的内容要求相对应。因为，同一内容领域的不同主题，或同一内容主题下的不同知识点，它们的重要性并不完全相同。一个内容主题或知识点越重要，或越具有包容性，其权重应该越大，测量这些内容主题或知识点的试题也就应该越多。

以上就是学业考试内容选择或内容改革必须遵循的四项原则。遗憾的是，目前各省市学业水平考试对内容的选择很少或几乎没有同时考虑这四项原则，其结果只能是：考试的结果并不能有效地反映考生的学业水平，学业水平考试也不可能与课程标准保持高度一致。

3. 试题编制和组卷未能遵循基本的教育测量原理

教育考试的根本目的是推测考生的心理结构。对于大规模的高中学业考试而言，这种心理结构就是考生经过中学阶段的学习后，所应具备的学科素养或学科能力，即其所掌握的知识以及对知识的建构能力。考生的学科素养或学科能力是抽象的，不可直接测量，我们必须通过可观测的变量对其加以标定。通过可观测的变量标定学生的学科能力或素养，一个基本的假设是：在不同观测变量上的学生观测值，完全是由其学科素养或学科能力所具有的量决定的。根据这个假设，学生所具有的学

科素养或学科能力的量，可以通过学生在不同观测变量上的观测值进行推断。学生在观测变量上的观测值可以通过观察学生对试题所给予的测量刺激的反应而获得。因此，从某种程度上来说，学业考试的本质，是通过考生在不同试题上的反应，推测其所具有的学科素养或学科能力的程度。

基于上述原理，学业水平考试命题的基本任务是：

（1）确定考试所要推测的心理结构，并确定对学科素养或学科能力推测所需要的观测变量，即确定与学业标准相对应的测量目标以及每个目标相应的行为目标或认知要求；

（2）编制试题，试题必须对考生产生合理的测量刺激，并就考生对试题作出的反应进行科学、合理的计分，即命制试题，制作科学、合理的评分量表，从而使考生在测量变量上的观测值变成其学科素养或学科能力的指标；

（3）组成合理的试卷，试卷必须从整体上考虑学业标准的要求，包括测量目标和每个目标相应的行为目标或认知要求的比例，考虑每个内容领域或主题知识的分配、比例和权重，考虑分类决策误差对不同难度、识别度试题比例和权重的要求。

目前的问题是，我国各省市学业水平考试命题几乎没有考虑这些基本任务和要求，突出表现在以下三个方面。(1)试题编制时，不考虑测量目标及其行为目标或认知要求；(2)制定评分量表未曾考虑测量目标及其行为目标或认知要求，评分的依据不是行为目标或认知要求，也没有按照学科能力的表现程度赋分。其结果只能是学生在试题上的得分与行为目标或认知要求没有关联，分数不具有可解释性；(3)组卷时，主要考虑的是内容领域或主题的知识的分配，而没有以学业标准为依据，没有考虑测量目标和每个目标相应的行为目标或认知要求的比例，更没有考虑分类决策误差以及分类决策误差对不同难度、识别度试题的比例和权重要求。

这样的学业考试结果，难以作为评价学生学业水平的依据，不能为分类决策提供科学的、高质量的数据。

4. 分类决策的过程欠科学

在我国，依据学业水平考试对学生进行分类决策，传统的是以 60 分为划界分数，小于 60 分为不合格，大于 60 分为合格。至于 60 分以下的内涵究竟是什么，60 分以下为什么就不合格，60 分以下的问题在哪里，谁也说不清。如果需要作出更多的分类，如优秀、良好、合格和不合格，或者 A、B、C、D 和 E 等等第，也总是以百分位分数作为分类的依据。前已述及，以百分位分数为依据，实际上是常模参照考试的做法，其本质上是将学生的表现与其他学生比较，与基于标准的教育体系宗旨不相符。

现代教育测量学要求，根据学业水平考试结果对学生进行分类决策，需要采用“标准设置”的过程。“标准设置”是按照规定的过程确定表现类型或学业水平边界（即划界分数）的活动。之所以要按照规定的过程，就是要保证“标准设置”确定的划

界分数有效，经得起检验。这个规定的过程包括：①选择标准设置的方法；②准备或熟悉表现类型(Performance Category)或学业水平描述；③组成“标准设置”专家小组；④对参与“标准设置”的专家进行培训；⑤向“标准设置”专家提供相关的反馈信息；⑥评价和记录过程的有效性。按照这个规定的过程，“标准设置”专家小组的任务是要依据自己的专业判断，将学业水平的定性描述，转换为考试的连续分数量表上具体的划界分数的位置。

尽管，有些学者对“标准设置”过程的主观判断性质提出批评，但“标准设置”过程已经成了教育测量界公认的对学生进行分类决策所必须的过程。

“标准设置”过程的测量学特征可以保证分类决策依据高质量的数据，并保证数据是以系统的、可重复的、客观的、可检验的方式组合和呈现。从更高的视野来看，如果必须对学生作出分类决策，那么采用规定的、系统的过程产生划界分数，并作出的分类决策，比人为的、以内涵不明确的60分或其他分数为标准作出的决策，会更公平、更明智、更有效、更具可检验性。

5. 分数报道不能提供有用信息

分数报道是我国教育考试最为薄弱的环节。各省市的学业水平考试通常是通知学校学生获得的等第，没有纸质或电子版的分数报道；有些省市虽然有纸质或电子版的分数报道，也只是每门学科的成绩及总分，或每门课的等第，没有任何关于学生学习和学校教学、课程绩效的信息。

笔者以为，学业水平考试及其成绩报道至少应该起到两个方面的作用：第一，成绩报道应该为学生的学习和学校的教学服务，应该向学生、家长、学校提供关于学生学习的信息、学校教学绩效的信息、课程对学生发展的作用和价值的信息；第二，成绩报道应该向不同层次的管理机构提供学生的学习成果以及整个教育体系的绩效等信息。

如果考虑学业水平考试成绩报道这两个方面的基本功能，那么，成绩报道至少应该包括以下要素。

(1) 报告对象

成绩报道的对象不应该是单一的，至少应该包括学生和家长、学校教师和校长、市区县教育行政部门，甚至省级教育行政部门。对不同的对象，报告的内容、方式应有所不同，但均应该以简明扼要的方式让报告对象清楚、全面地了解学生学习、教学和课程的绩效等信息。

(2) 成绩报道的量表

除等第外，我们要确定是报道原始分数还是量表分数。

(3) 学业标准的描述

学业水平考试的基本目的是评价学生在经过一定学段学习后，学业所达到的水

平，这种水平通常以等第表示，每一级等第的学生掌握了哪些知识、获得了哪些技能、能够做什么，即每个等第的内涵，需要通过学业标准的描述，传达给报告的对象。这是学业水平考试分数报道中非常核心的部分，它使得分数报道，不单纯是向报道对象报道等第或分数，而是使其理解学生的知识、技能和能力所达到的水平。此外，通过公开高水平的学业标准，可以促进学生和教学向高水平的标准努力，达到提高学生学习水平、教师教学质量的效果，这也是学校教育的重要目的之一。

(4) 评价单位

分数报道可以以试题为单位，也可以以大题或分测试为单位，还可以报道整个考试的总成绩。大题或分测试通常以内容领域、内容主题为单位，或者以相同的认知类型或技能为单位。

(5) 报道单位

最基本的报道单位是学生，但根据目的和报告对象的不同，报道单位也可以是班级、学校，甚至整个市、区、县。报道单位不同，报道的内容、信息、方式也有所不同。

如果报告对象是学校校长或教育行政部门，分数报道中通常应该包括测量误差。测量误差可以以总分作为划界分数处的条件标准测量误差，也可以报道分类决策的可靠性和准确性，以及它们的获取方法。如果分数报告以内容领域或认知领域等为评价单位，也应尽可能报道评价单位的条件标准测量误差或分类决策的可靠性和准确性。

(6) 分数的呈现方式

分数和等第的呈现可以是图表形式，也可以是文字描述形式，或者是两者、三者相结合的形式。学业水平考试结果的呈现方式应该根据对象的不同而有所变化，原则是尽可能简明扼要、清晰、准确地提供学习、教学的绩效信息。

(7) 报道媒介

分数报道可以采用纸质，也可以基于网络向不同的报告对象传递学业水平的信息。

总之，学业水平考试的分数报道只有从整体上考虑到这八个方面的因素，才能够向不同的对象提供有效的关于学生个人和群体的学业水平的信息。

二、学业水平考试问题的主要原因

上述学业水平考试五大方面的问题，使考试结果的有效性大打折扣。这些问题的出现和长期存在，与我国的教育管理机制、教育考试机构发展的历史和现状、教育考试理论和技术的落后等密切相关。

1. 现行的教育管理机制制约了考试的专业化发展

我国是一个考试大国，各类教育考试、职业能力考试、公务员选拔考试林林种种，

其基本特点是以政府为主导。就学业水平考试而言，都是以各省、市教育行政部门为主导的。教育行政部门不但从政策上规定了学业考试的目的、用途，还对学业考试的许多技术要求作了规定，如考试的分值、时间、题型、题量、难度、等第划分方式，等等。专业的考试机构无法在考试的设计、开发等方面取得话语权，只能按照教育行政部门文件的要求进行操作。在教育考试日益专业化的时代，政府完全控制学业考试政策制订和技术要求，甚至将教育考试机构科学的考试开发视为异端，不可避免地制约了考试的科学发展，降低了考试结果的有效性。

从这个角度来看，我国教育行政机制需要改革，教育行政机构只能从政策层面确定学业水平考试的目标和要求，而学业水平考试的设计、开发、管理等专业问题，应该交由专业的考试机构独立地进行。只有这样，教育行政机构依据由此获得的结果进行教育分类决策，才会更加科学、更加有效，决策的权威性才会大幅度地提高。

2. 我国没有教育和心理测量标准

学业水平考试在本质上属于教育和心理测量。教育和心理测量是一门理论和技术要求都很高的科学，如果不按照教育和心理测量的理论和技术要求进行考试开发，考试结果的有效性会大打折扣，根据考试结果作出的推断亦可能失效。为此，世界上教育和心理测量研究发达的国家都制订了教育和心理测量标准，如美国教育研究协会、美国心理学会和美国教育测量全国委员会早在 20 世纪 50 年代就联合制订了教育和心理测量标准(Standards for Educational and Psychological Testing)，该标准随着教育和测量理论及技术的发展不断修改、完善，成为美国教育和心理测量的国家标准。许多国家的教育和心理测量同行也以其为标准进行考试开发工作。美国的许多考试机构也依据此标准，制定了考试机构的内部标准，这种内部标准一般比国家标准更明确、更具体，对考试的设计、开发、管理的质量要求更高。目前，由于我国没有这些标准，有些教育行政部门任意地将学业水平考试交由任何与教育考试不相关的部门操作，且不管这个部门是否有能力从事学业考试的设计、开发与管理。许多人认为，无论考试是否满足教育和心理测量的要求，只要试题没有科学性、思想政治性问题，考试就是好考试。正因为没有必需的测量标准，第三方个人或机构也就无法对考试作出评价。

从教育考试事业发展的角度来看，我国需要加快制定教育和心理测量的国家标准，任何从事教育考试的机构和部门都必须按照标准的要求，从事学业水平的设计、开发和管理，对教育考试的评价也应以此为标准。惟有这样，才能保证我国的教育考试，包括学业水平考试，沿着科学的轨道前行。

3. 教育考试机构的专业化程度较低

这是我国所有教育考试(包括学业水平考试在内的)科学性、结果的有效性偏低的直接原因。

我国教育考试机构发展的历史普遍较短，早的也只有20年左右，多数考试机构只有几年的历史。早期的从业人员受教育程度普遍偏低。2004年以来，由于高考分省命题机制的建立，许多教育考试机构招收了一批硕士、博士或者高校和中学具有高级职称的教师，然而这些人主要是学科的专业人员，很少有人接受过教育和心理测量的专门学习和培训。因此，教育考试机构从业人员的专业化程度普遍偏低。其直接结果是，多数教育考试机构不能够依据教育测量的理论和技术要求独立地设计、开发、管理大规模的教育考试，包括学业水平考试。这是到目前为止，我国大多数学业考试问题多多、考试结果解释和使用的有效性不高的直接原因。

要提高教育考试的质量，迫切需要提高教育考试机构的专业化水平。首先，教育考试机构的领导要树立现代考试质量的理念，改变重招生，轻考试开发、命题的思想，注重考试的质量管理。其次，考试机构的从业人员必须提高专业化水平，要接受教育测量方面的培训，要自觉地以教育测量学的理论指导自己的考试实践，运用教育测量的方法和技术设计、开发和管理考试。

4. 课程标准与学业水平考试的要求不相适应

学业水平考试是基于标准的教育体系的一个重要环节，应该与课程标准、教学等调整一致。相应地，标准的制定也需要为考试评价创造有利条件。

西方基于标准的教育改革的核心思想是：设置明确的、可以测量的标准，对学生的学业表现提出高要求或高期望，课程、评价和教师的职业发展与标准调整一致（American Federation of Teachers，2003）。按照这样的核心思想，各国在推进教育改革的过程中，特别注重标准的制定。这里所说的标准通常包括三种类型：内容标准、课程标准和表现标准，表现标准又称为学业标准。

内容标准描述的是学生应该知道什么、能够做什么，或者应该掌握何种知识与技能。内容标准通常按照年级或年级段，以简明扼要的方式列出需要学习或教学的概念、原理、问题，以及相应的认知要求或认知技能，包括学科的一些专门技能——思维技能、操作技能、调查研究技能和交际技能。课程标准描述的是如何进行课堂教学，它关注的是教学方法、策略，以及推荐的教学活动。表现标准或学业标准描述的是学生学到多好才算好，即它定性地描述期望学生达到的学业水平。

西方国家一般颁布的是内容标准和表现标准，两者相伴出现，而课程标准相对较少。我国颁发的课程标准是一种更加广义的课程标准，多数糅合了西方国家内容标准和课程标准的内容，有些科目比较接近西方的纯课程标准，如语文学科的课程标准。这种课程标准的结构反映了我国教育行政机构关注课堂教学策略、方法和过程的质量观和价值观。这本身无可厚非，但教育部颁发的全国义务教育阶段课程标准和高中课程标准，以及上海市颁发的上海市中小学课程标准，均没有包含表现标准。这对学业水平考试的开发十分不利。

教育部已经下文要求各省、市建立与新课改相配套的学业水平考试，国家教育中长期发展规划，也要求各省市建立基础教育学业水平考试评价体系。这就要求各省市在国家颁发的新课程内容标准的基础上，根据本省基础教育发展的实际情况，设置学业标准。或者在教育部的主持下，对现行的课程标准进行修改，增加学业标准。惟此，学业水平考试才有科学发展的前提条件。

5. 教育和心理测量方面的教育水平偏低

目前，我国很少有高校开设教育和心理测量专业的课程，师范类院校虽然开设有教育评价课程，但教育评价课程与教育、心理测量还不完全相同，真正接受教育和心理测量教育的本科生、研究生并不多，即使是有些教育管理、师范类学生接受了教育评价课程的教育，也多数是重理论、轻实践，重理论、轻技术。加之，教育和心理测量不但对学生的文科素养要求较高，对学生的理科素养要求也较高，很多本科生和研究生，实际上并没有有效掌握教育和心理测量的理论和技术。这使得我国的教育管理和师范类本科、研究生的教育测量观念普遍淡漠，理论、技术落后。加之我国的考试实践与西方的考试实践不完全相同，一些本科生和研究生即使在学校学了一些教育和心理测量的知识，也难以与我国的考试实践相适应。难以形成有效的团队，难以担当教育考试设计、开发、管理的重任。这是长期以来我国教育考试事业落后、问题迟迟得不到解决的深层次原因。教育和心理测量教育的落后，必然导致教育考试事业的长期落后。

正是由于这些方面的原因，我国到目前为止还没有一本立足于我国的考试实践，系统阐述基于标准的教育考试的理论和实践的书籍。

作者正是基于这些思考和认识，撰写了本书。

本书共十五章。

第一章主要介绍基于标准的教育考试与常模参照考试的区别，基于标准的教育考试产生的国际背景及其特征。

第二章介绍基于标准的教育考试的效度检验的最新理念、方法和要求，以及这些理念、方法和要求对基于标准的教育考试的设计、开发的启示。许多人认为考试的效度检验是考试结束以后的事情，其实考试的效度检验在考试的设计阶段就需要开始了，效度检验所要求的各个方面的证据积累，贯穿于整个考试的设计、开发、施考、评分、考试结果处理、标准设置、成绩报告的全过程，仅仅到考试结束后的评价阶段再考虑效度检验，晚矣。

第三章介绍大规模基于标准的教育考试需要考虑的基本问题。笔者在《大规模教育考试：命题和评价》中也专辟一章讨论过大规模教育考试需要考虑的基本问题，不过那本书主要讨论的是大规模常模参照的教育考试设计的问题。而本书第三章侧

重在大规模基于标准的教育考试的设计问题，突出了大规模基于标准的教育考试设计的独特特征，而将两者的共同特征部分留待读者比较归纳。

第四章介绍大规模基于标准的教育考试中学业标准的开发。由于我国的课程标准中只有内容标准，缺少学业标准，任何从事大规模基于标准的教育考试的机构都必须依据课标的内容标准开发学业标准，否则就无法就进行大规模基于标准的教育考试的设计、开发和评价，也无法对考生进行学业水平标定。

第五章到第八章主要介绍基于标准的教育考试的试题命制。这四章针对我国课程内容标准的标准条目动词内涵不清的弊端，以 Bloom 的教育目标分类中的认知类型为框架，依据这些认知类型的内涵，介绍考查各种认知技能的试题撰写需要遵循的原则和要素。

第九章介绍主观题评分标准的开发。我国主观题命题的最大问题在于评分标准不能比较好地反映考生的认知技能或思维能力表现程度的差异。很多命题人员不清楚评分标准应该与试题的测量目标或试题考查的认知技能一致，应该按照考生认知技能或思维能力表现的程度赋分。因此，本章重点通过多学科评分标准开发的实例，介绍主观题评分标准开发的准则和方法。本章内容不但适用于基于标准的教育考试，也适用于常模参照考试中主观题评分标准的开发。

第十章讨论大规模基于标准的教育考试的内容选择和组卷要求。对于基于标准的大规模教育考试而言，其内容选择或组卷必须保证其考查的认知结构与内容标准的认知结构保持一致，知识内容的结构与内容标准的内容结构保持一致，认知技能和知识深度应该能够有效地区分不同学业水平的考生，保证依据考试分数所做出的分类决策具有足够高的分类一致性和分类准确性。

第十一章到第十三章介绍标准设置。其中第十一章介绍标准设置的概念，进行标准设置的技术过程和要求。第十二章介绍最常见的标准设置方法——Angoff 法。第十三章介绍最广泛使用的标准设置方法——Bookmark 法，即书签法。

第十四章介绍大规模基于标准的教育考试的信度。大规模基于标准的教育考试的信度概念和计算方法与大规模常模参照考试的信度概念和计算方法显著不同。教育考试机构的一些从业人员，尤其是学科秘书，多数都知道 Alpha 信度系数，尽管能够进行 Alpha 信度系数计算的不是太多。但知道分类一致性和分类准确性作为基于标准的大规模教育考试的信度的人可能少之又少，会进行实际计算的人可能更少。本章从传统的统计方法和项目反应理论方法两个方面介绍分类一致性和分类准确性的概念以及进行计算的程序和实例。

第十五章介绍大规模基于标准的教育考试的考试分数报告。如果说考试大纲是考试机构与考生、学校等交流考试信息的工具，那么考试分数报告是考试机构与考生、学校、教育行政机构交流考试结果，检查考生学习状况，诊断考生学习的强项和弱