# 浅谈美国国家教育进步评估中的科学评估

**武荷岚 [1,2]、杨友源 [1]、郑美红 [1]**

**1.香港教育学院数社科技学系**
**2.华东师范大学物理系**

**电邮：helanwu@126.com**

收稿日期：二零零八年一月二十日(于六月二十六日再修定)

## 内容



## 摘要

本文介绍了最近十年美国国家教育进步评估（NAEP）中的科学评估（1996、2000、2005年），对其评估框架和评分标准作了重点解读，并详细点评了这三次科学评估的例题，最后还就其启示作了简要的分析。

**关键词：**美国国家教育进步评估 NAEP；科学评估；学生学业成就

## 引言

美国国家教育进步评估（National Assessment of Educational Progress，简称 NAEP，「国进评」）也称为美国国家教育报告卡（Nation's Report Card TM），是目前美国唯一从全国范围内收集典型学生样本，且持续时间长达数十年的学生学业成绩评估体系。

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

本文主要介绍了最近十年「国进评」对学生科学学业水平的评估，由于最近十年的三次科学评估（1996 年、2000 年、2005 年）是在共同的评价框架下进行的，使得这三次评估不但具有纵向可比性，而且，代表着今后美国科学评估的趋势,成为今后科学评估的基础。故有必要对这三次科学评估做一个简要介绍，以便我们更好地理解其估目标和评分标准。

# 「国进评」的产生背景

1963 年，美国由于缺乏有关学生学业成绩方面的信息，国家教育专员弗兰西斯·凯普尔(Francis Keppel)呼吁建立一个全国性的学生学业成绩评估体系，并邀请著名的心理学家、教育家泰勒(Ralph W. Tyler) 共同参与筹备工作。由于是针对多个学科领域、多个年龄段，反映不同学生学业成就水平的全新评估方式，整个评估体系的开发时间比预期要长。到 1969 年，整个项目被重新命名为国家教育进步评估（NAEP）。

尽管从六十年代开始，「国进评」 就陆续在阅读、数学、科学、写作、历史、公民、地理和艺术等各个学科开展定期的学业成绩测评，测评对象是全美最具有代表性的 4 年级、8 年级以及 12 年级学生。但直至 1996 年，「国进评」 的评估模式的才完全确立：包括全国评估（National NAEP）、州评估（State NAEP）、城市地区试验性评估（NAEP Trial Urban District Assessment）、全国长期趋势评估（Long-Term Trend）等几类(周红, 2005)。下表是这几类评估的比较。

**表 1**：各种类型的「国进评」评估

| | **主要的全国评估 (Main National NAEP)** | **全国长期趋势评估 (Long-Term Trend)** | **州评估 (State NAEP)** | **实验性城市地区评估(TUDA)** |
|---|---|---|---|---|
| 评估科目 | 阅读、数学、科学、写作、历史、地理等各科目 | 阅读、数学 | 阅读、数学、(科学、写作是自选科目) | 阅读、数学、科学、写作 |
| 评估对象 | 4 年级、8 年级、12 年级 | 9 岁、13 岁、17 岁 | 4 年级和 8 年级(12 年级自选) | 一般为 4 年级和 8 年级公立学校的学生 |
| 评估时间 | 两年一次(每次 2-3 科) | 四年一次 | 两年一次(逢单年) | 两年一次(逢单年) |
| 评分方法 | 标准分数和等级水平 | 标准分数 | 标准分数和等级水平 | 标准分数和等级水平 |

每次「国进评」完成后，最终的结果以学生的性别、种族、学校类型、所在地区、背景信息变量等类别进行报告，不报告参与评估的学生或学校的个别信息。参与的各州可以将评

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

价结果与全国或其它州的学生平均水平相比较，与本州岛的目标相比较,明确学生学业所在的全国水平，发现本州岛在教育上的不足，为改进教育工作提供参考。

# 「国进评」的科学评估框架

「国进评」科学评估是在名为「评估框架」(Framework)的蓝图指导下进行的(Loomis & Bourque, 2001)，1996 年，「国进评」科学评估推出了 1996-2000 新的科学评估框架，该框架是美国国家评估管理委员会（NAGB）组织学科专家、科学家、学校行政人员、决策者、教师、家长等共同开发而成的，它规定了应该如何评定 4、8、12 这三个年级学生的学业水平。由于该评估框架是在前面 1990 评估的基础上建立起来的，故具有一定的包容性(Stiggins, 1987)。

框架分为科学领域(Fields of Science)与认知要素(Elements of Knowing and Doing Science)两个纬度（如下图）。科学领域涉及地球科学、物质科学和生命科学，其中物质科学包括物理和化学；认知要素则细分为概念理解(Conceptual Understanding)、科学探究(Scientific Investigation)和实用推理(Practical Reasoning)三个要素(Allen & Zelenak, 1999)。

下图是科学评估框架，而图内列出一些典型的例子：

| | | 科学领域 | | |
|---|---|---|---|---|
| | | 地球科学 | 物质科学 | 生命科学 |
| 认知要素 | 概念理解 | 例题 4：风蚀现象 | | 例题 1：辨认身体器官的功能 |
| | 科学探究 | | 例题 3：判断盐水和纯水<br>例题 5：判断金属的纯度 | |
| | 实用推理 | 例题 6：地球轨道 | 例题 2：判断容积大小 | |

注：另外还有科学本质、专题研习两个维度，它们一般不被单独列出，而是渗透在前面的要素中考察。

在该评估框架的认知领域中，概念理解要素重点考察学生对科学知识和概念的理解，其中，科学知识包括从学校科学教育以及自然界中学习到的各种事实、事件，以及用于解释、预测自然现象的科学概念、定律和理论。科学探究主要考查学生使用科学工具的能力，包括制定计划，使用多种科学工具获得信息，交流探究的结果等。实际推理则考查学生在新的、真实世界中运用其科学理解能力。它们在最近十年三次测评中所占比例参见表 3。

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

表 2：认知要素纬度

| 认知要素 | 说明 |
|---|---|
| 概念理解(Conceptual Understanding) | 学生对用于解释和预测自然世界中各种现象的科学概念和原理的理解； |
| 科学探究(Scientific Investigation) | 运用科学知识和技能，设计合适的调查计划和步骤，利用各种科学工具探寻新知。 |
| 实用推理 (Practical Reasoning) | 运用科学知识来解决日常生活的问题。 |

表 3：认知要素在各年级中的比例分布

| 认知要素 | 4 年级 | | | 8 年级 | | | 12 年级 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 |
| 概念理解 | 45% | 56% | 50% | 45% | 59% | 55% | 44% | 56% | 56% |
| 科学探究 | 38% | 27% | 40% | 29% | 18% | 29% | 28% | 24% | 29% |
| 实用推理 | 17% | 17% | 10% | 26% | 24% | 16% | 28% | 20% | 14% |

从表 3 可以看出，「国进评」科学评价非常强调对科学概念的理解，它在各个年级所占百分比都几乎达到 50%左右，2005 年 12 年级最高，达到 56%；同时还可以看出，年级越低，越强调科学探究（4 年级 2005 年最高占 40%，12 年级 2000 年最低占 24%），说明，对于年龄较小的学生，在“做中学”的学习方法越为重要。另外，随着年级增加，对学生实际应用能力的要求也在逐步的提高，1996 年 4 年级占 17%，12 年级推理能力要求占到 28%。这也是符合学生成长、认知规律的。可是，近年(2005 年)的趋势显示实用推理的比重明显地下降了。

评估框架内的各认知要素一般是相互联系的，因为反映任何概念的重要性不仅来自于它自身的事实和观点，而且来自于与之相联系的方法和技能，即在要求学生掌握科学的事实和观点的同时，要求他们应用科学概念去理解和探究，从而建构逻辑的推理方法。

各科学领域包涵的内容和主题见下表 4。最近十年的 3 次科学评估中，各科学领域在三个年级中比例分布见下表 5。

表 4：各科学领域包涵的内容

| 科学领域 | 包涵的内容 |
|---|---|
| 地球科学 | 地球科学包括地壳（土壤和岩石圈）、水（水蒸气）、空气（大气）和地球空间等通常见的主题。 |
| 物质科学 | 物质科学包括物理和化学，涵盖了关于宇宙结构以及有关物质运作原理的基 |

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

| | |
|---|---|
| | 本知识和理解。主要议题是物质及其转变、能量和它的转换、物体的运动。 |
| 生命科学 | 生命科学的主要目标是理解和解释自然和生命系统的性质和功能。主要概念有生物的变化与进化，细胞及其功能（四年级不包括此项），有机体，生态学等。 |

表 5：科学领域在各年级中的比例分布

| 科学领域 | 4 年级 | | | 8 年级 | | | 12 年级 | | |
|---|---|---|---|---|---|---|---|---|---|
| | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 | 1996 | 2000 | 2005 |
| 地球科学 | 33% | 34% | 33% | 30% | 31% | 30% | 33% | 33% | 34% |
| 物质科学 | 34% | 33% | 34% | 30% | 34% | 33% | 33% | 30% | 35% |
| 生命科学 | 33% | 33% | 33% | 40% | 35% | 37% | 34% | 37% | 31% |

（资料来源：美国国家教育统计中心，「国进评」 1996、2000 和 2005 年科学测评）

从以上表 5 可以看出，最近十年评估中，地球科学在 4、8、12 年级中所占比例几乎未曾改变，一直约占总分的三分之一。生命科学在 8 年级的比例稍偏高，超过 35%；2005 年，物质科学在各年级中的比例都有了不同程度的提高。

# 目标分类依据和评分标准

## 3.1 「国进评」科学评估的目标分类依据

1993 年，美国科学促进会（简称 AAAS）出版了《科学素养的基准》，「国进评」科学评估的目标分类是依据其制定的以探究学习为核心的(Science-A Process Approach）科学过程技能训练目标。AAAS 从科学家对自己科研活动的大体描述中抽取了 13 种科学方法或过程作为测评的目标，分别为 8 种属基本技能:观察、运用时空关系、分类、数字应用、测量、交流、预测、推理，及 5 种综合技能: 解释数据、控制变量、建立假设、操作定义、实验(Frank ,1957; Haertel, 1991; Davis, 1990）。

「国进评」还对上述 13 种过程技能进行了详细的描述性解释。并将这些科学过程技能的要求以附录形式附在中学理科教科书中，以此作为标准来设计教科书中的习题，还在每一习题前注明检测的是何种技能，以使学生能针对自己在某方面技能或某些能力的欠缺，调整自己的学习，也有助于教师了解学生对某种技能的掌握情况，做到因材施教，有利于学生科学探究能力的培养与发展。

## 3.2 「国进评」科学评估的评分标准

「国进评」评分方法有两种，一种是标准分数（scale scores），阅读、数学、历史、地理各科总分是 0-500 分，科学、写作、公民学各科总分是 0-300 分；另一种是等级水平

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

（achievement level），分为基本合格（Basic）、熟练（Proficient）和优秀（Advanced）三个等级。

在「国进评」 的科学学业成就评估中，两种评分方法同时使用，既有标准分数，又有等级水平(Bourque, Champagne & Crissman, 1997)。科学问卷总分是 300 分，对应着四个等级，“基本合格”水平表明部分掌握该年级的基本知识与概念；“良好”水平表明具有扎实的学术表现，学生达到这个层次的水平，可以完成该年级具有挑战性的科学问题；“优秀”等级则是具有出众的学术表现，如果低于基本合格水平就是不合格等级。各个等级在各年级对应的标准分值见下表 6：

**表 6**：等级水平和标准分数对应表

| 等级水平 | 标准的界定 | 4 年级 | 8 年级 | 12 年级 |
|---|---|---|---|---|
| 基本合格 **(Basic)** | 这个等级水平是指学生的知识和技能达到该等级的基本部分。 | 138 | 143 | 146 |
| 良好 **(Proficient)** | 学生达到这个水平，代表着稳定的学业表现。能对较具有挑战性的问题进行研究，包括应用学到的知识，分析真实世界的情况，并有适当的解决问题的技巧。 | 170 | 170 | 178 |
| 优秀 **(Advanced)** | 达到这个水平的学生表现很优异。 | 205 | 208 | 210 |

（资料来源：美国国家评估委员会 2000）

各年级在 3 个等级相对应的标准分数见上表，后一个等级水平高于前一个等级水平。相同等级和分数，不同年级对应的要求不同，例如都是基本水平，4 年级 138 分是基本合格水平，8 年级是 143 分，12 年级需要 146 分（见表 6）；再如 4 年级的熟练水平和 8 年级熟练水平尽管都是要求 170 分，但是具体内容和要求显然也会不同(Aldridge,1989, Bybee,1989)。

另外，「国进评」还用一张图表（Science Item Map）来说明各年级 0-300 分对应的具体问题和对学生具体技能的要求(Resnick, 1987)。例如「国进评」 4 年级评分标准表（见表 7）。

**表 7**：「国进评」评分标准表（4 年级）

| 等级水平 | 分值 | 问题描述的例子 |
|---|---|---|
| 优秀（205） | 219 | 理解雨水测量仪的读数。 |
| | 208 | 解读图表数据来总结种子发芽所需的条件。 |
| 良好 | 203 | 解释从化石中可以获取的信息。 |

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

| | | |
|---|---|---|
| （170） | 185 | 找出空气（氧气）与燃烧时间的关系。 |
| | 174 | 根据熔点资料，判断哪种东西先熔化。 |
| 基本合格（138） | 165 | 根据图表，判断哪一天的日照时间最长。 |
| | 159 | 预测和解释两件物体的排水量。 |
| | 139 | 确认某人体结构的功能。 |
| 不合格 | 136 | 辨认鱼儿获得氧气的过程。 |
| | 103 | 根据气象数据比较不同城市的温度高低。 |

如何正确理解表中问题与对应的标准分值间的关系，具体介绍请见后面第六部分的例题。

# 测量工具的设计及评估过程

「国进评」科学评估之所以能评判和比较全国各州学生的学业水平，除了在相同的评价框架指导下，使用相同的评估程序外，还因为整个评估使用了有效的评估工具：大型题库和矩阵技术，并在不同年份的测试、不同年级的测试卷中特意安排了一些重迭的问题(Shymansky et al.,1983; Stiggins,1987; Strang, 1990)，与其它类型的测试共享部分相同的样本，使得测试结果既有纵向可比性又有横向可比性。

下面简要介绍「国进评」科学评估的问卷组成和题库设计以及评估过程。

## 4.1 问卷组成

「国进评」科学评估量表由学生问卷（测试卷、学生背景调查问卷）、教师问卷、学校领导问卷等组成。

学生问卷分为两部分：第一部分是有关科学学科内容的主题模块，每份问卷中通常刊载三个模块，每个主题模块包括十几个评估框架中提到的有关认知和技能的问题。它们被随机放置在学生问卷的小册子里，同一所学校的学生接到的问卷题目可能并不相同。

另一部分是有关背景资料的问题，问题包括学生的种族、父母受教育情况、家庭经济状况，就读学校的类型（公立还是私立），是否接受语言辅助，有没有享受免费午饭计划等一些被认为对学生学习情况有影响的相关信息。

除了学生问卷，整个评估还设置了对教师和学校管理者的问卷，和学校纪录卡，以作为背景资料的重要来源。教师问卷、学校管理者问卷需要参与评估的学校教师、领导用几分钟的时间填写自己学校情况，如：学校性质，学生种族比例，是否有残疾学生，对残疾学生是否有辅助以及测试完成时间等。

## 4.2 题库的设计

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

「国进评」科学评估除了与别的测试（如州「国进评」评估）共享部分相同的样本外，「国进评」科学评估还使用了大型题库，例如，在 2000 年的总题库中，4、8、12 年级分别有 143 题、196 题、195 题。

总题库的题目围绕一定的主题分成一些模块，每个年级题库中通常有 15 个主题模块，每个参加测试的学生只需要回答 3 个这样的主题模块。这样既可节省学生参加测试的时间，测试题目又可以包含足够宽广的知识。

题型包括选择题、简答题和问答题三种。简答题(Short constructed-response questions )通常需要一两个句子来回答（例如，简单地说明为什么盆栽植物能够比老鼠更长时间地生存在一个密封的货柜里），问答题(extended constructed response questions)需要用一整段语句来回答（例如，概述测量金属环密度的实验工具和步骤）。问答题往往具有拓展性，包括几个问题，有时答案并不唯一，有的需要图解、图表，或计算等。

**表 8**：2000 年和 1996 年的题库中题型分布表

| 年级 | 选择题 | | 简答题 | | 问答题 | |
|---|---|---|---|---|---|---|
| | 1996 | 2000 | 1996 | 2000 | 1996 | 2000 |
| **4 年级** | 51 | 71 | 73 | 65 | 16 | 7 |
| **8 年级** | 74 | 95 | 100 | 91 | 20 | 10 |
| **12 年级** | 70 | 91 | 88 | 83 | 30 | 21 |

此外，每个参加评估的学校中有半数以上的学生必须完成一道实验操作题。实验操作题往往给学生一套设备，让其进行探究，并在答题纸上回答相关的问题。例如，8 年级学生有可能被要求，基于提供的有关太阳系的数据，画出图纸和图表，然后回答与该主题相关的一些问题；又如给 12 年级的学生一瓶新的饮料，它被认为是无糖和无卡路里的，问学生该如何判断情况是否属实。学生需要设计实验步骤，选取和列出需要的器材，动手实验，记录下实验数据，并解释得出结论的推理过程或依据。

「国进评」在设计题库的时候，特意在不同年份、不同年级的测试卷中安排了一些相同或重迭(overlap)的问题，具体说明如下见表 9：

a) 不同年份的测试，相同年级的试题有部分重迭。b) 相同年份的测试，不同年级的试题也有部分重迭。

**表 9**：重迭题目的题型分布表（1996，2000 年科学测试）

| | 选择题 | 简答题 | 问答题 | 总数目 |
|---|---|---|---|---|
| **4 年级和 8 年级** | 9 | 16 | 4 | 29 |
| **8 年级和 12 年级** | 21 | 26 | 3 | 50 |

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

所谓矩阵技术，就是指每个参加测评的学生只需要完成整个题库的一小部分，最后整合在一起，通过矩阵运算就可以算出该生的总成绩、所有参加测试学生的平均成绩和对应的等级水平。

正是通过题库和矩阵技术，既能评估学生对科学概念的理解、运用高层思维的能力与技巧，又使得测试结果具有纵、横向可比性。

### 4.3 选样和施测

在取样方面，「国进评」不是完全随机抽样，而是在参加评估的州内根据人口统计学和地理组成进行抽样。并且为了保证样本的均衡性，NCES 和 NAGB 规定州和地方学校的参与率不得低于 85%。通常在各州 4 年级和 8 年级各选取 100 所学校作为样本，再在作为样本的学校和年级选取 25 名学生参加每个科目的评估[5]。2005 年，有 44 个州 30 多万名学生参加了科学评估，其中，4 年级和 8 年级学生参与率达到 85%，12 年级略微低于这个数字。

测试时间方面，「国进评」制定了详细的评估计划表。全国评估通常与州评估和实验性城市地区评估是来年进行并且避免重迭，而全国长期趋势评估则是四年一次。下次全国长期趋势评估是 2008 年，但从 2002 年开始，停止了对科学科的长期趋势评估。

答题时间，4 年级的学生必须在 20 分钟内完成一个主题模块（通常包括 2-3 个主题模块），实验操作题必须在 20 分钟内完成。8 年级和 12 年级的学生分别在 25 分钟和 30 分钟内完成一个主题模块；一半学生需另外在 30 分钟内完成一个给定的实验操作任务，并且回答与任务有关的问题。加上回答背景问题的 10 分钟，故 4 年级总答题时间是 70 分钟，8、12 年级总答题时间是 100 分钟。

评估进行时，每个模块单独计时。即当一个主题模块的时间用尽时，会有工作人员通知答题时间到，所有学生停止回答该模块，然后再进行下一个模块的答题。

## 最近十年的 3 次「国进评」科学评估结果

我们可以从政府公布的 3 份评估报告（又称国家教育报告卡）中了解最近十年的 3 次「国进评」科学评估的情况和结果。2005 年，「国进评」对美国 30 多万学生进行了科学评估，2000 年对 2,078 所学校的 93,993 学生，1996 年对 4,812 所学校的 23,000 学生进行了科学评估。报告卡对全国学生的学业表现作了报告，并比较了 3 次测试的结果，并对 4、8 年级的结果和教育体验、学校环境等方面的信息加以了适当说明。

### 5.1 全国学生的平均成绩

如表 10 所示，2005 年评估结果总体情况是低年级学生比高年级学生进步明显。2005 年 4 年级学生均分为 151 分，2000 年和 1996 年均分 147 分，有显著上升，并且达到基本合格水平的学生占 68%，比起 1996 年（63%）和 2000 年（63%）都有了明显进步；2005 年 12 年级学生总体水平下降，尽管与 2000 年（146 分）没有太大变化，但都低于 1996 年 150 分的成绩。有 54%的学生及格，18%的 12 年级学生达到良好水平，他们知道燃

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

烧反应中的热能转换。最近十年中 8 年级学生的科学成绩没有提高，及格率几乎都是 59%，其中有 29%的学生达到良好及以上水平(参照表 6 中关于标准及等级的界定)。

**表 10**：全国学生在三次科学评估中的平均成绩和及格率

| | 1996 年 | | 2000 年 | | 2005 年 | |
|---|---|---|---|---|---|---|
| | 均分 | 及格率 | 均分 | 及格率 | 均分 | 及格率 |
| **4 年级** | 147* | 63%* | 147* | 63%* | 151 | 68% |
| **8 年级** | 149 | 60% | 149 | 59% | 149 | 59% |
| **12 年级** | 150* | 57%* | 146 | 52% | 147 | 54% |

注：*表示差异显著(与 2005 年比较)。

资料来源：美国教育部，教育科学学院，国家教育进步评估，1996, 2000，2005 科学评估 http://nces.ed.gov/nationsreportcard/itmrls/

### 5.2 全国学生在三门科学科目中的具体表现

由表 11 可看到 8 年级学生的地球科学成绩在 2005 年有了显著提高，比 2000 年 150 分上升 3 分；物质科学成绩明显下降，从 1996 年的 150 分下降到 146 分；生命科学没有明显变化；12 年级成绩明显降低，2005 年与 2000 年没有太大变化，但都比 1996 年成绩明显降低，其中物质科学更是明显，从 1996 年的 150 分下降到 2005 年 145 分。

**表 11**：8、12 年级学生在三门科学科目中的具体表现

| | 8 年级 | | | 12 年级 | | |
|---|---|---|---|---|---|---|
| | 1996 年 | 2000 年 | 2005 年 | 1996 年 | 2000 年 | 2005 年 |
| **地球科学** | 149 | 150 | 153* | 151* | 145 | 145 |
| **物质科学** | 150* | 148* | 146 | 150* | 147 | 145 |
| **生命科学** | 149 | 150 | 150 | 150* | 148 | 148 |

（注：*表示差异显著(与 2005 年比较)。表中只列出 8 年级和 12 年级的结果，由于 4 年级还没有分科，故无法列出。）

资料来源：美国教育部，教育科学学院，国家教育进步评估，1996, 2000，2005 科学评估 http://nces.ed.gov/nationsreportcard/itmrls/

### 5.3 性别差异

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

全国结果中存在性别差异，10 年的测试结果表明，男生在三门科学科目中的成绩普遍好于女生，但男女学生在不同年份和不同年级的表现有着不同。4 年级全体学生 2005 年成绩比 2000 年有了显著上升，其中男生均分达到 153 分，比女生 149 分高出 4 分。12 年级的男女生成绩普遍下降，并且男生成绩下降更大，从 1996 年的 154 分下降到 2005 年 149 分，男女生差异从 1996 年相差 7 分到 2005 年相差 4 分；8 年级男女生差异从 2000 年相差 7 分到 2005 年相差 3 分，故可以认为女生进步比男生快，男女生成绩差异正在缩小。

**表 12**：3 次测试男女学生的平均成绩

| | | 1996 年 | **2000 年** | **2005 年** |
|---|---|---|---|---|
| **4 年级** | 男生 | 148* | 149* | 153 |
| | 女生 | 146 | 145* | 149 |
| **8 年级** | 男生 | 150 | 153* | 150 |
| | 女生 | 148 | 146 | 147 |
| **12 年级** | 男生 | 154* | 148 | 149 |
| | 女生 | 147* | 145 | 145 |

注：*表示差异显著(与 2005 年比较)。

资料来源：美国教育部，教育科学学院，国家教育进步评估，1996, 2000，2005 科学评估

### 5.4 社会经济因素

另外，「国进评」的调查报告还对学校类型、所在地区、学生家庭特征，小区和学校对学生成绩的影响，不同的学习机会及教育政策对学生学业成绩影响等信息变量作了研究和分析，结果如下：

#### a)父母受教育程度对学生学业的影响

「国进评」科学学业调查发现：总体来说，父母受教育程度与学生的学业表现成正相关，父母受教育程度高的孩子学业表现好。这与别的调查结果相一致。2005 年评估发现，美国约二分之一的 8 年级和 12 年级学生的父母中至少有一人是大学毕业，只有 6%的学生父母高中未毕业。比较后还发现 12 年级学生学业表现与父母的受教育程度相关度下降，其中原因尚不明确，一种原因可能是随着学生年岁的增加，学习的内驱力在增加，受外界影响要变小。8 年级则较为明显，见下表。

**表 13**：8 年级学生学业表现与父母的教育程度对应表（2005）

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

| 父母受教育程度 | 高中未毕业或以下 | 高中毕业 | 大专毕业 | 大学毕业 | 不知道 |
|---|---|---|---|---|---|
| 学生平均成绩 | 128 | 138 | 151 | 159 | 130 |

资料来源：美国教育部，教育科学学院，国家教育进步评估，1996, 2000，2005 科学评估 http://nces.ed.gov/nationsreportcard/itmrls/

**b)不同种族学生的学业表现存在差异**

以 2005 年 8 年级学生成绩为例，白人学生平均得分为 160，亚洲学生为 156，西班牙语的学生为 129 而黑人学生则为 124。与 2000 年比较，4、8 年级的少数民族学生进步明显，4 年级黑人学生(占人口总数的 16%)，与西班牙裔学生(占人口总数的 19%)都取得显著进步。

**c)不同类型学校的学生学业表现存在差异**

公立学校和私立学校的学生表现明显不同，私立学校的学生学业表现明显要好于公立学校学生(Braun et al., 2006)，以 8 年级为例，见表 14。

**表 14**：8 年级学生的学业成绩分类表

| 学校类型 | 1996 | 2000 | 2005 | 午餐是否获资助 | 1996 | 2000 | 2005 |
|---|---|---|---|---|---|---|---|
| 公立学校 | 148 | 148 | 147 | 是 | 129 | 127 | 130 |
| 其它私立学校 | 165 | 167 | 未知 | 否 | 156 | 159 | 159 |
| 天主教资助的学校 | 161 | 165 | 163 | 未知 | 157 | 155 | 160 |

资料来源：美国教育部，教育科学学院，国家教育进步评估，1996, 2000，2005 科学评估 http://nces.ed.gov/nationsreportcard/itmrls/

**d)不同地段学生的学业表现存在差异**

调查中把学校划分为中心城市学校、郊区学校、乡村学校三种类型，结果发现，中心城市学生学业成绩最差，乡村学生成绩最好。8 年级中心城市均分 141 分，比乡村学校的学生（152 分）均分低 11 分。但 12 年级差异没有这么明显。

**f) 家庭收入与学生的学业表现有显著的相关性**

由国家农业部资助的学校免费(或减费)午餐计划，让家庭收入低于规定标准的学生可以接受资助。调查结果发现这部分学生的学业表现显著地低于未接受资助的学生成绩，但他们比 2000 年成绩有了明显进步（以 8 年级为例，见表 14）。

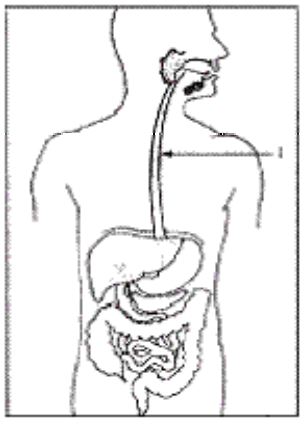
Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

此外，「国进评」还对各州的结果进行了比较，找出进步最快的州，并研究其教育政策。

## 「国进评」试题举例及分析

「国进评」每次只公开其试题的一小部分，所以有时无法了解每次评估的全貌，现从「国进评」已经公开的部分试题，举例分析如下。

**例题 1**：（4 年级）



上图中有人体内的一些器官。请问箭头 1 所指的器官的主要功能是什么？

A 运输空气 B 通过食物 C 运输血液 D 传递大脑发出的信息

答案：B

[点评]这是一道 4 年级有关生物科学的选择题，主要考察学生辨认不同组织、器官的功能。本题对应 4 年级的 140 分的要求，属于基本难度的题目，分数达到 140 分以上的学生中，有超过 74%的学生能正确回答本题。

**例题 2**：（4 年级）一个炎热的日子里，你将要去一个公园玩，需要随身多带一些水解渴。假设你有下图中所示的三个盛水瓶子，你想要带盛水最多的水瓶。你请问，你如何确定哪个是你想要的？

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

样题提供的答案：你可以用几个杯子，分别往瓶里面灌水，灌的杯数最多的水瓶就是最大的，是需要带走的。

[点评]本题是 4 年级的一道简答题，考查学生如何判断哪个容器的容积大。需要动脑分析并思考如何操作。问题具有情境性，让学生在解决生活问题的过程中运用所学到的科学知识。本题相当于 4 年级 226 分的要求，属于优秀等级题。在 4 年级分数达到 226 分的学生中，有 65%以上的学生能基本正确回答本题。

**例题 3：**（8 年级）

玛丽亚有两杯水，其中一杯是纯净水，另外一杯是盐水。请问，她该如何判断哪杯是盐水？（不用品尝）

[点评]这是一道 8 年级的有关物理知识的问答题，让学生判断哪个是盐水。本题相当于 8 年级 230 分的要求，试题具有开放性，给学生自由发挥的空间。

**例题 4：**

（8 年级） 古埃及艳后克利欧佩特拉的石针是竖立在埃及沙漠数千年的一块大石碑。自从被搬到纽约市中心公园几年后，它的表面就开始剥落。

(1)请问是什么原因导致其剥落？

(2)纽约市希望其能继续留在纽约市中心公园，请问如何才能防止它进一步恶化？

（注：本题评分按照错误(Incorrect)，部分正确(Partial)，正确(Complete)三个等级来评定。）

参考答案：

(1)因为污染和酸雨导致。

(2)他们可以用屋顶或者别的东西罩住它，使得它免遭酸雨的破坏。

[点评]本题是 8 年级的一道考察地球科学知识的简答题，要求学生会分析和说明石碑风化的原因，并想出办法来阻止进一步恶化。问题具有情境性，让学生在真实的问题解决中运用所学到的科学知识。本题属于 8 年级的基本题，相当于 8 年级 144 分的要求。在 8 年级分数达到 144 分的美国学生中，有超过 65%的学生能基本正确回答本题。

**例题 5：**（12 年级）

金属的密度是可以用来区分不同金属的一种重要性质，你可以根据下表来判断是何种金

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

属，

| 金属 | 金 | 铅 | 银 | 铜 | 锡 | |
|---|---|---|---|---|---|---|
| 密度（$g/cm^3$） | 19.3 | 11.3 | 10.5 | 8.9 | 7.3 | |

假如给你一枚戒指，请你判断它是否是纯金做的。请设计你判断其密度的步骤。并解释如何操作，使用的器材等。

（注：本题评分按照错误(Incorrect)，部分正确(Partial)，正确(Complete)三个等级来评定。）

[点评]本题是 12 年级考察物理实验操作的问答题。需要学生自行设计实验步骤，自己考虑如何使用器材，步骤不限，具有一定的开放性。本题相当于 12 年级 194 分的要求。在 12 年级分数达到 194 分的美国学生中，有超过 65%的学生能正确回答本题。

**例题 6**：（12 年级）

用工具在地面上观测太阳，发现 1 月份的太阳会比 7 月份的太阳稍微大些，请从下面选出正确的原因。（）

A 地球绕着太阳沿着椭圆轨道运转，在 1 月份比在 7 月份更靠近太阳。

B 地球的直径不是常数，在赤道上会大些，并且冬季也是如此。

C 地球的轨道不像其它行星在同一个平面。

D 地球的旋转轴不是垂直于轨道平面，而是倾斜一个角度。

答案：A

[点评]本题是有关地球科学的 12 年级的一道选择题。本题属于 12 年级的优秀等级的题目，相当于 12 年级 223 分的要求。在 12 年级分数达到 223 分的美国学生中，有超过 74%的学生能正确回答本题。

从以上 6 道例题可以看出，「国进评」的测试题并不难，但有一个很重要的特点就是，无论是选择题还是问答题，都是尽量从学生身边的问题入手，考查学生在具体的问题情境中解决问题的能力。并且问答题答案并不唯一，往往具有一定的开放性。这就避免学生死记硬背一些知识,这一点显然很有意义。当然，也有些题目的性质是近似硬背知识，例如：例题 1.

# 评论及总结

作为全美唯一从全国不同地区，从三个不同年龄段采集学业成绩的信息体系，「国进评」持续时间长达数十年，其样本数量远大于美国其它一些代表性的学生成绩调查，如美国

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

有名的教育纵向调查（NESL）样本容量仅占「国进评」的 10%-20%。并且「国进评」样本包含不同种族的学生和学生的家庭、教育背景等资料。它能基本反映美国初等和中等科学教育的现状，能揭示出美国各州、各地区在教育工作上的不足，为改进教育工作提供参考，是当之无愧地 “美国国家教育的晴雨表”。

除此之外，我们认为最近十年的 3 次「国进评」科学评估，还具有以下几个特点值得关注。

（1）「国进评」科学评估强调概念理解；从「国进评」的评估框架和测试题可以看出，概念理解在各个年级均占 45%以上，「国进评」科学评估尤其强调那些学生对身边的科学概念的理解(Resnick, 1987)，突出科学概念在日常生活情景中的运用，而不仅仅是考核学生对概念的识记(苑大勇, 2007)。

（2）「国进评」科学评估强调科学探究的过程；「国进评」科学评价强调概念理解的同时还非常强调科学探究，把科学探究作为学生获取知识的工具(朱行建,2007; 赵保钢,2007)。并且年级越低，越强调科学探究（4 年级最高占 38%，12 年级最低占 28%）。希望学生在一系列探究的过程中，用亲身的体验和行为来加深对科学概念的理解，并在探究的过程中将科学的概念应用在新的情况中，以便更好地实现知识的迁移。

（3）「国进评」科学评估重视科学方法的掌握；「国进评」的评估框架列出的一些探究性的科学主题中，除了需要学生寻找数据和数据，确定、筛选有用的信息外，「国进评」科学评估还强调学生的解读和使用图表的能力。因为能否高效地获取信息，是学生未来能否很好适应信息爆炸的社会的重要方面。例如 4 年级的评分标准图中就多次提到根据图表数据判断、解释问题，问答题和实验题更是经常需要学生通过图表有效的表达结果。

（4）「国进评」科学评估强调对科学知识应用技能的真实性评估；「国进评」科学评估考核学生能力比较全面、合理。它一改美国传统的客观性试题一统天下的局面，采用了选择题、问答题和实验操作题。用选择题测评学生对重要事实和概念的掌握情况，用问答题重点考察学生理解、分析、应用、传达科技信息的能力，采用实验操作题，真实性地评价学生观察、动手实验等能力。既有纸笔测验，也有实验动手操作，对难以考察的探究能力、理论运用于实践的真实水平是一个很好的途径。

（5）注重考察学生的思维过程，尤其是学生的高层思维能力；「国进评」科学评估中的不少试题具有开放题的性质，评估测试的过程中，不强调答案的唯一性，而是尽可能发散，尽可能让学生思考，并给其选择的权利，选择合适的测量工具，设计自己的实验步骤。重点查看学生有没有掌握科学的思维过程，用科学的思维在相同或不同的领域里处理新的问题。鼓励学生自己建立、测试和修改理论模型，提倡学生独立思考，引发学生的高层次思维。

同时，「国进评」试题的设计，按照项目反应理论，采用了大型题库和矩阵技术，既尽可能全面地考核学生所学，又不致增加学生的测试负担，使得评估更加高效。

当然，「国进评」科学评估也有其自身的不足：如评估结果发布的时间间隔太长（18—24 个月），评估标准还可以进一步改进。同时，由于「国进评」本身的设计模式决定了仅

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

能收集 4、8、12 这三个年级的学校教学实践的同期信息，却无法知道学生在别的学年的学习体验（例如学生前一个学年已有的学习经验），使得其比较基点不完全相同，这是它无法克服的弊端。

从美国「国进评」的结果来看，多项社会经济因素，包括性别、父母教育程度、种族、学校类型、地区、家庭收入等都会与学生科学成绩之间存在明显的相关性。此外，从该三次「国进评」成绩的纵向比较，我们也可看到美国科学教育的多个层面都在不断地改进中。以上的社会经济因素引致的差异也有不少改善的趋势。

对于这些结果和当中的原因是否适用于中国内地的科学教育情况，都是值得我们(包括一线科学教师、教育官员和科学教育学者/研究员)参考、反思及/或作出深入的进一步研究。但是，在采纳或引用美国「国进评」的结果或评估方法前，我们必须小心谨慎地考虑当中的社会文化的差异、教育制度和课程上的差异，评估目的以及评估结果的不同和可能用途。

综上所述，「国进评」科学评估的经验和结果告诉人们：好的教育评估应该既能起到全面了解教育现状，检验教学效果的作用，又能充分发挥正面导向作用，引导科学教育：重视学生对科学概念的理解，提倡探究式的科学学习，增加学生亲身体验，从而真正把所学运用于日常生活的问题解决中。让科学不再是抽象的学术知识，让学生灵活地掌握新经济要求的知识和技能，以便学生更好地适应未来社会。

# 参考文献

Aldridge, B.G. (1989). *Essential Changes in Secondary School Science: Scope, Sequence and Coordination*. Washington, DC: National Science Teachers Association.

Allen, N. L., Carlson, J., & Zelenak, C.A. (1999) *The NAEP 1996 technical report*. Washington, DC: National Center for Education Statistics. [Online] http://nces.ed.gov/nationsreportcard/itmrls/

American Association for the Advancement of Science (1989). *Science for All Americans: A Project 2061 Report on Literacy Goals in Science, Mathematics, and Technology*. Washington, DC: American Association for the Advancement of Science.

American Geological Institute (1991). *Earth Science Education for the 21st Century: A Planning Guide*. Alexandria, VA: American Geological Institute.

Bourque, M. L., Champagne, A. B., & Crissman, S. (1997). *1996 science performance standards: Achievement results for the nation and the states*. Washington, DC: National Assessment Governing Board.

Braun, H., Jenkins, F., and Grigg, W. (2006). *Comparing Private Schools and Public Schools Using Hierarchical Linear Modeling (NCES 2006-461)*. U. S. Department of Education, National Center for Education Statistics, Institute of Education Sciences. Washington, DC: U.S. Government Printing Office. [Online] http://nces.ed.gov/nationsreportcard/science/distributequest.asp

Bybee, R.W., C.E. Buchwald, S. Crissman, D.R. Heil, P.J. Kuerbis, C. Matsumoto, and J.D. McInerney (1989). *Science and Technology Education for the Elementary Years: Frameworks for Curriculum and Instruction*. Washington, DC: National Center for Improving Science Education.

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.

California Department of Education (1990). *Science Framework for California Public Schools: Kindergarten Through Grade Twelve*. Sacramento, CA: California Department of Education.

Davis, F. (1990). Assessing Science Education: A Case for Multiple Perspectives. In G.E. Hein (Ed.), *The Assessment of Hands-On Elementary Science Programs*. Grand Forks, ND: North Dakota Study Group.

Frank, P. (1957). *Philosophy of Science, the Link Between Science and Philosophy*. A Spectrum Book. Englewood Cliffs, NJ: Prentice-Hall.

Haertel, E. (1991). R*eport on TRP analyses of issues concerning within-age versus cross-age scales for the National Assessment of Educational Progress*. Washington, DC: National Center for Education Statistics. (ERIC Document Reproduction Service No. 404367)

Loomis, S. C., & Bourque, M. L. (Eds.). (2001d). *National Assessment of Educational Progress achievement levels, 1992–1998 for science*. Washington, DC: National Assessment Governing Board.

National Center for Education Statistics. (n.d.). *Percentage of vocational and nonvocational public school teachers of grades 9 to 12 by selected demographic and educational characteristics: 1999- 2000* [EB/ OL]. [Online] http://nces.ed.gov/nationsreportcard/nde/, 2006-12-24.

National Center for Improving Science Education (1990). *Science and Technology Education for the Middle Years: Frameworks for Curriculum and Instruction*. Washington, DC: The National Center for Improving Science Education.

National Center for Improving Science Education (1991). *The High Stakes of High School Science*. Washington, DC: The National Center for Improving Science Education.

National Research Council (1990). *Fulfilling the Promise: Biology Education in the Nation's Schools.* Washington, DC: National Academy Press.

National Science Board Commission on Precollege Education in Mathematics, Science and Technology (1983). *Educating Americans for the 21st Century*. Washington, DC: National Science Foundation.

Resnick, L. (1987). *Education and Learning to Think*. Washington, DC: National Academy Press.

Shymansky, J.A., W.C. Kyle, and J.M. Alpert (1983). The Effects of New Science Curricula on Student Performance. *Journal of Research in Science Teaching*, 20(5): 387 04.

Stiggins, R.J. (1987). Design and Development of Performance Assessment. *Educational Measurement*: Issues and Practices, Fall: 33 2.

Strang, J. (1990). *Measurement in School Science*. London, United Kingdom: Assessment Performance Unit, School Examination and Assessment Council, Central Office of Information.

周红，美国国家教育进展评估 NAEP 体系的产生与发展，外国教育研究，2005 年第 2 期。

苑大勇，美国《2005 年城市地区学生科学能力试验性评估》报告简述，教育研究，2007 年 3 月。

朱行建，国际教育评价中的科学探究能力测评简介及启示，课程教材教法，2007 年 2 月。

赵保钢，美国全国教育进步评价（NAEP）中的科学探究，物理教学探讨，2005 年第 7 期(上半月)。

Copyright (C) 2008 HKIEd APFSLT. Volume 9, Issue 1, Article 17 (Jun., 2008). All Rights Reserved.