淺談美國國家教育進步評估中的科學評估

武荷嵐 [1,2]、楊友源 [1]、鄭美紅 [1]

1.香港教育學院數社科技學系
2.華東師範大學物理系

電郵：helanwu@126.com

收稿日期：二零零八年一月二十日(於六月二十六日再修定)

內容

摘要

本文介紹了最近十年美國國家教育進步評估（NAEP）中的科學評估（1996、2000、2005年），對其評估框架和評分標準作了重點解讀，並詳細點評了這三次科學評估的例題，最後還就其啓示作了簡要的分析。

關鍵字: 美國國家教育進步評估 NAEP；科學評估；學生學業成就

引言

美國國家教育進步評估（National Assessment of Educational Progress，簡稱 NAEP，「國進評」）也稱爲美國國家教育報告卡（Nation's Report Card TM），是目前美國唯一從全國範圍內收集典型學生樣本，且持續時間長達數十年的學生學業成績評估體系。

本文主要介紹了最近十年「國進評」對學生科學學業水準的評估，由於最近十年的三次科學評估（1996 年、2000 年、2005 年）是在共同的評價框架下進行的，使得這三次評估不但具有縱向可比性，而且，代表著今後美國科學評估的趨勢,成爲今後科學評估的基礎。故有必要對這三次科學評估做一個簡要介紹，以便我們更好地理解其估目標和評分標準。

「國進評」的產生背景

1963 年，美國由於缺乏有關學生學業成績方面的資訊，國家教育專員弗蘭西斯・凱普爾(Francis Keppel)呼籲建立一個全國性的學生學業成績評估體系，並邀請著名的心理學家、教育家泰勒(Ralph W. Tyler) 共同參與籌備工作。由於是針對多個學科領域、多個年齡段，反映不同學生學業成就水準的全新評估方式，整個評估體系的開發時間比預期要長。到 1969 年，整個項目被重新命名爲國家教育進步評估（NAEP）。

儘管從六十年代開始，「國進評」 就陸續在閱讀、數學、科學、寫作、歷史、公民、地理和藝術等各個學科開展定期的學業成績測評，測評物件是全美最具有代表性的 4 年級、8 年級以及 12 年級學生。但直至 1996 年，「國進評」 的評估模式的才完全確立：包括全國評估（National NAEP）、州評估（State NAEP）、城市地區試驗性評估（NAEP Trial Urban District Assessment）、全國長期趨勢評估（Long-Term Trend）等幾類(周紅, 2005)。下表是這幾類評估的比較。

表 1：各種類型的「國進評」評估

	主要的全國評估 (Main National NAEP)	**全國長期趨勢評估 (Long-Term Trend)**	**州評估 (State NAEP)**	**實驗性城市地區評估(TUDA)**
評估科目	閱讀、數學、科學、寫作、歷史、地理等各科目	閱讀、數學	閱讀、數學、(科學、寫作是自選科目)	閱讀、數學、科學、寫作
評估對象	4 年級、8 年級、12 年級	9 歲、13 歲、17 歲	4 年級和 8 年級(12 年級自選)	一般爲 4 年級和 8 年級公立學校的學生
評估時間	兩年一次(每次 2-3 科)	四年一次	兩年一次(逢單年)	兩年一次(逢單年)
評分方法	標準分數和等級水準	標準分數	標準分數和等級水準	標準分數和等級水準

每次「國進評」完成後，最終的結果以學生的性別、種族、學校類型、所在地區、背景資訊變數等類別進行報告，不報告參與評估的學生或學校的個別資訊。參與的各州可以將評

價結果與全國或其他州的學生平均水準相比較，與本州的目標相比較，明確學生學業所在的全國水準，發現本州在教育上的不足，爲改進教育工作提供參考。

「國進評」的科學評估框架

「國進評」科學評估是在名爲「評估框架」(Framework)的藍圖指導下進行的(Loomis & Bourque, 2001)，1996 年，「國進評」科學評估推出了 1996-2000 新的科學評估框架，該框架是美國國家評估管理委員會（NAGB）組織學科專家、科學家、學校行政人員、決策者、教師、家長等共同開發而成的，它規定了應該如何評定 4、8、12 這三個年級學生的學業水準。由於該評估框架是在前面 1990 評估的基礎上建立起來的，故具有一定的包容性(Stiggins, 1987)。

框架分爲科學領域(Fields of Science)與認知要素(Elements of Knowing and Doing Science)兩個緯度（如下圖）。科學領域涉及地球科學、物質科學和生命科學，其中物質科學包括物理和化學；認知要素則細分爲概念理解(Conceptual Understanding)、科學探究(Scientific Investigation)和實用推理(Practical Reasoning)三個要素(Allen & Zelenak, 1999)。

下圖是科學評估框架，而圖內列出一些典型的例子：

		科學領域		
		地球科學	物質科學	生命科學
認知要素	概念理解	例題 4：風蝕現象		例題 1：辨認身體器官的功能
	科學探究		例題 3：判斷鹽水和純水 例題 5：判斷金屬的純度	
	實用推理	例題 6：地球軌道	例題 2：判斷容積大小	

注：另外還有科學本質、專題研習兩個維度，它們一般不被單獨列出，而是滲透在前面的要素中考察。

在該評估框架的認知領域中，概念理解要素重點考察學生對科學知識和概念的理解，其中，科學知識包括從學校科學教育以及自然界中學習到的各種事實、事件，以及用於解釋、預測自然現象的科學概念、定律和理論。科學探究主要考查學生使用科學工具的能力，包括制定計劃，使用多種科學工具獲得資訊，交流探究的結果等。實際推理則考查學生在新的、真實世界中運用其科學理解能力。它們在最近十年三次測評中所占比例參見表 3。

表 2：認知要素緯度

認知要素	說明
概念理解(Conceptual Understanding)	學生對用於解釋和預測自然世界中各種現象的科學概念和原理的理解；
科學探究(Scientific Investigation)	運用科學知識和技能，設計合適的調查計畫和步驟，利用各種科學工具探尋新知。
實用推理 (Practical Reasoning)	運用科學知識來解決日常生活的問題。

表 3：認知要素在各年級中的比例分佈

認知要素	4 年級			8 年級			12 年級		
	1996	2000	2005	1996	2000	2005	1996	2000	2005
概念理解	45%	56%	50%	45%	59%	55%	44%	56%	56%
科學探究	38%	27%	40%	29%	18%	29%	28%	24%	29%
實用推理	17%	17%	10%	26%	24%	16%	28%	20%	14%

從表 3 可以看出，「國進評」科學評價非常強調對科學概念的理解，它在各個年級所占百分比都幾乎達到 50%左右，2005 年 12 年級最高，達到 56%；同時還可以看出，年級越低，越強調科學探究（4 年級 2005 年最高占 40%，12 年級 2000 年最低占 24%），說明，對於年齡較小的學生，在“做中學”的學習方法越爲重要。另外，隨著年級增加，對學生實際應用能力的要求也在逐步的提高，1996 年 4 年級占 17%，12 年級推理能力要求占到 28%。這也是符合學生成長、認知規律的。可是，近年(2005 年)的趨勢顯示實用推理的比重明顯地下降了。

評估框架內的各認知要素一般是相互聯繫的，因爲反映任何概念的重要性不僅來自於它自身的事實和觀點，而且來自於與之相聯繫的方法和技能，即在要求學生掌握科學的事實和觀點的同時，要求他們應用科學概念去理解和探究，從而建構邏輯的推理方法。

各科學領域包涵的內容和主題見下表 4。最近十年的 3 次科學評估中，各科學領域在三個年級中比例分佈見下表 5。

表 4：各科學領域包涵的內容

科學領域	包涵的內容
地球科學	地球科學包括地殼（土壤和岩石圈）、水（水蒸氣）、空氣（大氣）和地球空間等通常見的主題。
物質科學	物質科學包括物理和化學，涵蓋了關於宇宙結構以及有關物質運作原理的基

	本知識和理解。主要議題是物質及其轉變、能量和它的轉換、物體的運動。
生命科學	生命科學的主要目標是理解和解釋自然和生命系統的性質和功能。主要概念有生物的變化與進化，細胞及其功能（四年級不包括此項），有機體，生態學等。

表 5：科學領域在各年級中的比例分佈

科學領域	4 年級			8 年級			12 年級		
	1996	2000	2005	1996	2000	2005	1996	2000	2005
地球科學	33%	34%	33%	30%	31%	30%	33%	33%	34%
物質科學	34%	33%	34%	30%	34%	33%	33%	30%	35%
生命科學	33%	33%	33%	40%	35%	37%	34%	37%	31%

（資料來源：美國國家教育統計中心，「國進評」 1996、2000 和 2005 年科學測評）

從以上表 5 可以看出，最近十年評估中，地球科學在 4、8、12 年級中所占比例幾乎未曾改變，一直約占總分的三分之一。生命科學在 8 年級的比例稍偏高，超過 35%；2005 年，物質科學在各年級中的比例都有了不同程度的提高。

目標分類依據和評分標準

3.1 「國進評」科學評估的目標分類依據

1993 年，美國科學促進會（簡稱 AAAS）出版了《科學素養的基準》，「國進評」科學評估的目標分類是依據其制定的以探究學習爲核心的(Science-A Process Approach）科學過程技能訓練目標。AAAS 從科學家對自己科研活動的大體描述中抽取了 13 種科學方法或過程作爲測評的目標，分別爲 8 種屬基本技能:觀察、運用時空關係、分類、數字應用、測量、交流、預測、推理，及 5 種綜合技能: 解釋資料、控制變數、建立假設、操作定義、實驗(Frank ,1957; Haertel, 1991; Davis, 1990）。

「國進評」還對上述 13 種過程技能進行了詳細的描述性解釋。並將這些科學過程技能的要求以附錄形式附在中學理科教科書中，以此作爲標準來設計教科書中的習題，還在每一習題前注明檢測的是何種技能，以使學生能針對自己在某方面技能或某些能力的欠缺，調整自己的學習，也有助於教師瞭解學生對某種技能的掌握情況，做到因材施教，有利於學生科學探究能力的培養與發展。

3.2 「國進評」科學評估的評分標準

「國進評」評分方法有兩種，一種是標準分數（scale scores），閱讀、數學、歷史、地理各科總分是 0-500 分，科學、寫作、公民學各科總分是 0-300 分；另一種是等級水準

（achievement level），分爲基本合格（Basic）、熟練（Proficient）和優秀（Advanced）三個等級。

在「國進評」 的科學學業成就評估中，兩種評分方法同時使用，既有標準分數，又有等級水準(Bourque, Champagne & Crissman, 1997)。科學問卷總分是 300 分，對應著四個等級，“基本合格”水準表明部分掌握該年級的基本知識與概念；“良好”水準表明具有扎實的學術表現，學生達到這個層次的水準，可以完成該年級具有挑戰性的科學問題；“優秀”等級則是具有出眾的學術表現，如果低於基本合格水準就是不合格等級。各個等級在各年級對應的標準分值見下表 6：

表 6：等級水準和標準分數對應表

等級水準	標準的界定	4 年級	8 年級	12 年級
基本合格 **(Basic)**	這個等級水準是指學生的知識和技能達到該等級的基本部分。	138	143	146
良好 **(Proficient)**	學生達到這個水準，代表著穩定的學業表現。能對較具有挑戰性的問題進行研究，包括應用學到的知識，分析真實世界的情況，並有適當的解決問題的技巧。	170	170	178
優秀 **(Advanced)**	達到這個水準的學生表現很優異。	205	208	210

（資料來源：美國國家評估委員會 2000）

各年級在 3 個等級相對應的標準分數見上表，後一個等級水準高於前一個等級水準。相同等級和分數，不同年級對應的要求不同，例如都是基本水準，4 年級 138 分是基本合格水準，8 年級是 143 分，12 年級需要 146 分（見表 6）；再如 4 年級的熟練水準和 8 年級熟練水準儘管都是要求 170 分，但是具體內容和要求顯然也會不同(Aldridge,1989, Bybee,1989)。

另外，「國進評」還用一張圖表（Science Item Map）來說明各年級 0-300 分對應的具體問題和對學生具體技能的要求(Resnick, 1987)。例如「國進評」 4 年級評分標準表（見表 7）。

表 7：「國進評」評分標準表（4 年級）

等級水準	分值	問題描述的例子
優秀（205）	219	理解雨水測量儀的讀數。
	208	解讀圖表資料來總結種子發芽所需的條件。
良好	203	解釋從化石中可以獲取的資訊。

（170）	185	找出空氣（氧氣）與燃燒時間的關係。
	174	根據熔點資料，判斷哪種東西先熔化。
基本合格（138）	165	根據圖表，判斷哪一天的日照時間最長。
	159	預測和解釋兩件物體的排水量。
	139	確認某人體結構的功能。
不合格	136	辨認魚兒獲得氧氣的過程。
	103	根據氣象資料比較不同城市的溫度高低。

如何正確理解表中問題與對應的標準分值間的關係，具體介紹請見後面第六部分的例題。

測量工具的設計及評估過程

「國進評」科學評估之所以能評判和比較全國各州學生的學業水準，除了在相同的評價框架指導下，使用相同的評估程式外，還因爲整個評估使用了有效的評估工具：大型題庫和矩陣技術，並在不同年份的測試、不同年級的測試卷中特意安排了一些重疊的問題(Shymansky et al.,1983; Stiggins,1987; Strang, 1990)，與其他類型的測試共用部分相同的樣本，使得測試結果既有縱向可比性又有橫向可比性。

下面簡要介紹「國進評」科學評估的問卷組成和題庫設計以及評估過程。

4.1 問卷組成

「國進評」科學評估量表由學生問卷（測試卷、學生背景調查問卷）、教師問卷、學校領導問卷等組成。

學生問卷分爲兩部分：第一部分是有關科學學科內容的主題模組，每份問卷中通常刊載三個模組，每個主題模組包括十幾個評估框架中提到的有關認知和技能的問題。它們被隨機放置在學生問卷的小冊子裏，同一所學校的學生接到的問卷題目可能並不相同。

另一部分是有關背景資料的問題，問題包括學生的種族、父母受教育情況、家庭經濟狀況，就讀學校的類型（公立還是私立），是否接受語言輔助，有沒有享受免費午飯計畫等一些被認爲對學生學習情況有影響的相關資訊。

除了學生問卷，整個評估還設置了對教師和學校管理者的問卷，和學校紀錄卡，以作爲背景資料的重要來源。教師問卷、學校管理者問卷需要參與評估的學校教師、領導用幾分鐘的時間填寫自己學校情況，如：學校性質，學生種族比例，是否有殘疾學生，對殘疾學生是否有輔助以及測試完成時間等。

4.2 題庫的設計

「國進評」科學評估除了與別的測試（如州「國進評」評估）共用部分相同的樣本外，「國進評」科學評估還使用了大型題庫，例如，在 2000 年的總題庫中，4、8、12 年級分別有 143 題、196 題、195 題。

總題庫的題目圍繞一定的主題分成一些模組，每個年級題庫中通常有 15 個主題模組，每個參加測試的學生只需要回答 3 個這樣的主題模組。這樣既可節省學生參加測試的時間，測試題目又可以包含足夠寬廣的知識。

題型包括選擇題、簡答題和問答題三種。簡答題(Short constructed-response questions)通常需要一兩個句子來回答（例如，簡單地說明爲什麼盆栽植物能夠比老鼠更長時間地生存在一個密封的貨櫃裏），問答題(extended constructed response questions)需要用一整段語句來回答（例如，概述測量金屬環密度的實驗工具和步驟）。問答題往往具有拓展性，包括幾個問題，有時答案並不唯一，有的需要圖解、圖表，或計算等。

表 8：2000 年和 1996 年的題庫中題型分佈表

年級	選擇題		簡答題		問答題	
	1996	2000	1996	2000	1996	2000
4 年級	51	71	73	65	16	7
8 年級	74	95	100	91	20	10
12 年級	70	91	88	83	30	21

此外，每個參加評估的學校中有半數以上的學生必須完成一道實驗操作題。實驗操作題往往給學生一套設備，讓其進行探究，並在答題紙上回答相關的問題。例如，8 年級學生有可能被要求，基於提供的有關太陽系的資料，畫出圖紙和圖表，然後回答與該主題相關的一些問題；又如給 12 年級的學生一瓶新的飲料，它被認爲是無糖和無卡路里的，問學生該如何判斷情況是否屬實。學生需要設計實驗步驟，選取和列出需要的器材，動手實驗，記錄下實驗資料，並解釋得出結論的推理過程或依據。

「國進評」在設計題庫的時候，特意在不同年份、不同年級的測試卷中安排了一些相同或重疊(overlap)的問題，具體說明如下見表 9：

a) 不同年份的測試，相同年級的試題有部分重疊。b) 相同年份的測試，不同年級的試題也有部分重疊。

表 9：重疊題目的題型分佈表（1996，2000 年科學測試）

	選擇題	簡答題	問答題	總數目
4 年級和 8 年級	9	16	4	29
8 年級和 12 年級	21	26	3	50

所謂矩陣技術，就是指每個參加測評的學生只需要完成整個題庫的一小部分，最後整合在一起，通過矩陣運算就可以算出該生的總成績、所有參加測試學生的平均成績和對應的等級水準。

正是通過題庫和矩陣技術，既能評估學生對科學概念的理解、運用高層思維的能力與技巧，又使得測試結果具有縱、橫向可比性。

4.3 選樣和施測

在取樣方面，「國進評」不是完全隨機抽樣，而是在參加評估的州內根據人口統計學和地理組成進行抽樣。並且爲了保證樣本的均衡性，NCES 和 NAGB 規定州和地方學校的參與率不得低於 85%。通常在各州 4 年級和 8 年級各選取 100 所學校作爲樣本，再在作爲樣本的學校和年級選取 25 名學生參加每個科目的評估[5]。2005 年，有 44 個州 30 多萬名學生參加了科學評估，其中，4 年級和 8 年級學生參與率達到 85%，12 年級略微低於這個數字。

測試時間方面，「國進評」制定了詳細的評估計畫表。全國評估通常與州評估和實驗性城市地區評估是隔年進行並且避免重迭，而全國長期趨勢評估則是四年一次。下次全國長期趨勢評估是 2008 年，但從 2002 年開始，停止了對科學科的長期趨勢評估。

答題時間，4 年級的學生必須在 20 分鐘內完成一個主題模組（通常包括 2-3 個主題模組），實驗操作題必須在 20 分鐘內完成。8 年級和 12 年級的學生分別在 25 分鐘和 30 分鐘內完成一個主題模組；一半學生需另外在 30 分鐘內完成一個給定的實驗操作任務，並且回答與任務有關的問題。加上回答背景問題的 10 分鐘，故 4 年級總答題時間是 70 分鐘，8、12 年級總答題時間是 100 分鐘。

評估進行時，每個模組單獨計時。即當一個主題模組的時間用盡時，會有工作人員通知答題時間到，所有學生停止回答該模組，然後再進行下一個模組的答題。

最近十年的 3 次「國進評」科學評估結果

我們可以從政府公佈的 3 份評估報告（又稱國家教育報告卡）中瞭解最近十年的 3 次「國進評」科學評估的情況和結果。2005 年，「國進評」對美國 30 多萬學生進行了科學評估，2000 年對 2,078 所學校的 93,993 學生，1996 年對 4,812 所學校的 23,000 學生進行了科學評估。報告卡對全國學生的學業表現作了報告，並比較了 3 次測試的結果，並對 4、8 年級的結果和教育體驗、學校環境等方面的資訊加以了適當說明。

5.1 全國學生的平均成績

如表 10 所示，2005 年評估結果總體情況是低年級學生比高年級學生進步明顯。2005 年 4 年級學生均分爲 151 分，2000 年和 1996 年均分 147 分，有顯著上升，並且達到基本合格水準的學生占 68%，比起 1996 年（63%）和 2000 年（63%）都有了明顯進步；2005 年 12 年級學生總體水準下降，儘管與 2000 年（146 分）沒有太大變化，但都低於 1996 年 150 分的成績。有 54%的學生及格，18%的 12 年級學生達到良好水準，他們知道燃

燒反應中的熱能轉換。最近十年中 8 年級學生的科學成績沒有提高，及格率幾乎都是 59%，其中有 29%的學生達到良好及以上水準(參照表 6 中關於標準及等級的界定)。

表 10：全國學生在三次科學評估中的平均成績和及格率

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

注：*表示差異顯著(與 2005 年比較)。

資料來源：美國教育部，教育科學學院，國家教育進步評估，1996, 2000，2005 科學評估 http://nces.ed.gov/nationsreportcard/itmrls/

5.2 全國學生在三門科學科目中的具體表現

由表 11 可看到 8 年級學生的地球科學成績在 2005 年有了顯著提高，比 2000 年 150 分上升 3 分；物質科學成績明顯下降，從 1996 年的 150 分下降到 146 分；生命科學沒有明顯變化；12 年級成績明顯降低，2005 年與 2000 年沒有太大變化，但都比 1996 年成績明顯降低，其中物質科學更是明顯，從 1996 年的 150 分下降到 2005 年 145 分。

表 11：8、12 年級學生在三門科學科目中的具體表現

	8 年級			12 年級		
	1996 年	2000 年	2005 年	1996 年	2000 年	2005 年
地球科學	149	150	153*	151*	145	145
物質科學	150*	148*	146	150*	147	145
生命科學	149	150	150	150*	148	148

（注：*表示差異顯著(與 2005 年比較)。表中只列出 8 年級和 12 年級的結果，由於 4 年級還沒有分科，故無法列出。）

資料來源：美國教育部，教育科學學院，國家教育進步評估，1996, 2000，2005 科學評估 http://nces.ed.gov/nationsreportcard/itmrls/

5.3 性別差異

全國結果中存在性別差異，10 年的測試結果表明，男生在三門科學科目中的成績普遍好于女生，但男女學生在不同年份和不同年級的表現有著不同。4 年級全體學生 2005 年成績比 2000 年有了顯著上升，其中男生均分達到 153 分，比女生 149 分高出 4 分。12 年級的男女生成績普遍下降，並且男生成績下降更大，從 1996 年的 154 分下降到 2005 年 149 分，男女生差異從 1996 年相差 7 分到 2005 年相差 4 分；8 年級男女生差異從 2000 年相差 7 分到 2005 年相差 3 分，故可以認爲女生進步比男生快，男女生成績差異正在縮小。

表 12：3 次測試男女學生的平均成績

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

注：*表示差異顯著(與 2005 年比較)。

資料來源：美國教育部，教育科學學院，國家教育進步評估，1996, 2000，2005 科學評估

5.4 社會經濟因素

另外，「國進評」的調查報告還對學校類型、所在地區、學生家庭特徵，社區和學校對學生成績的影響，不同的學習機會及教育政策對學生學業成績影響等資訊變數作了研究和分析，結果如下：

a)父母受教育程度對學生學業的影響

「國進評」科學學業調查發現：總體來說，父母受教育程度與學生的學業表現成正相關，父母受教育程度高的孩子學業表現好。這與別的調查結果相一致。2005 年評估發現，美國約二分之一的 8 年級和 12 年級學生的父母中至少有一人是大學畢業，只有 6%的學生父母高中未畢業。比較後還發現 12 年級學生學業表現與父母的受教育程度相關度下降，其中原因尚不明確，一種原因可能是隨著學生年歲的增加，學習的內驅力在增加，受外界影響要變小。8 年級則較爲明顯，見下表。

表 13：8 年級學生學業表現與父母的教育程度對應表（2005）

父母受教育程度	高中未畢業或以下	高中畢業	大專畢業	大學畢業	不知道
學生平均成績	128	138	151	159	130

資料來源：美國教育部，教育科學學院，國家教育進步評估，1996, 2000，2005 科學評估 http://nces.ed.gov/nationsreportcard/itmrls/

b)不同種族學生的學業表現存在差異

以 2005 年 8 年級學生成績爲例，白人學生平均得分爲 160，亞洲學生爲 156，西班牙語的學生爲 129 而黑人學生則爲 124。與 2000 年比較，4、8 年級的少數民族學生進步明顯，4 年級黑人學生(占人口總數的 16%)，與西班牙裔學生(占人口總數的 19%)都取得顯著進步。

c)不同類型學校的學生學業表現存在差異

公立學校和私立學校的學生表現明顯不同，私立學校的學生學業表現明顯要好于公立學校學生(Braun et al., 2006)，以 8 年級爲例，見表 14。

表 14：8 年級學生的學業成績分類表

學校類型	1996	2000	2005	**午餐是否獲資助**	1996	2000	2005
公立學校	148	148	147	是	129	127	130
其他私立學校	165	167	未知	否	156	159	159
天主教資助的學校	161	165	163	未知	157	155	160

資料來源：美國教育部，教育科學學院，國家教育進步評估，1996, 2000，2005 科學評估 http://nces.ed.gov/nationsreportcard/itmrls/

d)不同地段學生的學業表現存在差異

調查中把學校劃分爲中心城市學校、郊區學校、鄉村學校三種類型，結果發現，中心城市學生學業成績最差，鄉村學生成績最好。8 年級中心城市均分 141 分，比鄉村學校的學生（152 分）均分低 11 分。但 12 年級差異沒有這麼明顯。

f) 家庭收入與學生的學業表現有顯著的相關性

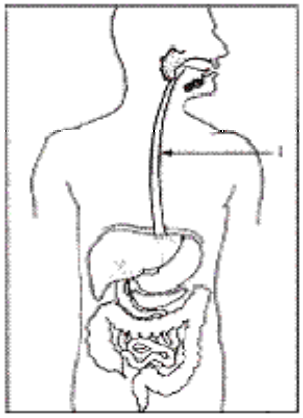
由國家農業部資助的學校免費(或減費)午餐計畫，讓家庭收入低於規定標準的學生可以接受資助。調查結果發現這部分學生的學業表現顯著地低於未接受資助的學生成績，但他們比 2000 年成績有了明顯進步（以 8 年級爲例，見表 14）。

此外，「國進評」還對各州的結果進行了比較，找出進步最快的州，並研究其教育政策。

「國進評」試題舉例及分析

「國進評」每次只公開其試題的一小部分，所以有時無法瞭解每次評估的全貌，現從「國進評」已經公開的部分試題，舉例分析如下。

例題 1：（4 年級）



上圖中有人體內的一些器官。請問箭頭 1 所指的器官的主要功能是什麼？

A 運輸空氣 B 通過食物 C 運輸血液 D 傳遞大腦發出的資訊

答案：B

[點評]這是一道 4 年級有關生物科學的選擇題，主要考察學生辨認不同組織、器官的功能。本題對應 4 年級的 140 分的要求，屬於基本難度的題目，分數達到 140 分以上的學生中，有超過 74%的學生能正確回答本題。

例題 2：（4 年級）一個炎熱的日子裏，你將要去一個公園玩，需要隨身多帶一些水解渴。假設你有下圖中所示的三個盛水瓶子，你想要帶盛水最多的水瓶。你請問，你如何確定哪個是你想要的？

樣題提供的答案：你可以用幾個杯子，分別往瓶裏面灌水，灌的杯數最多的水瓶就是最大的，是需要帶走的。

[點評]本題是 4 年級的一道簡答題，考查學生如何判斷哪個容器的容積大。需要動腦分析並思考如何操作。問題具有情境性，讓學生在解決生活問題的過程中運用所學到的科學知識。本題相當於 4 年級 226 分的要求，屬於優秀等級題。在 4 年級分數達到 226 分的學生中，有 65%以上的學生能基本正確回答本題。

例題 3：（8 年級）

瑪麗亞有兩杯水，其中一杯是純淨水，另外一杯是鹽水。請問，她該如何判斷哪杯是鹽水？（不用品嘗）

[點評]這是一道 8 年級的有關物理知識的問答題，讓學生判斷哪個是鹽水。本題相當於 8 年級 230 分的要求，試題具有開放性，給學生自由發揮的空間。

例題 4：

（8 年級） 古埃及豔後克利歐佩特拉的石針是豎立在埃及沙漠數千年的一塊大石碑。自從被搬到紐約市中心公園幾年後，它的表面就開始剝落。

(1)請問是什麼原因導致其剝落？

(2)紐約市希望其能繼續留在紐約市中心公園，請問如何才能防止它進一步惡化？

（注：本題評分按照錯誤(Incorrect)，部分正確(Partial)，正確(Complete)三個等級來評定。）

參考答案：

(1)因爲污染和酸雨導致。

(2)他們可以用屋頂或者別的東西罩住它，使得它免遭酸雨的破壞。

[點評]本題是 8 年級的一道考察地球科學知識的簡答題，要求學生會分析和說明石碑風化的原因，並想出辦法來阻止進一步惡化。問題具有情境性，讓學生在真實的問題解決中運用所學到的科學知識。本題屬於 8 年級的基本題，相當於 8 年級 144 分的要求。在 8 年級分數達到 144 分的美國學生中，有超過 65%的學生能基本正確回答本題。

例題 5：（12 年級）

金屬的密度是可以用來區分不同金屬的一種重要性質，你可以根據下表來判斷是何種金

屬，

金屬	金	鉛	銀	銅	錫	
密度（g/cm^3）	19.3	11.3	10.5	8.9	7.3	

假如給你一枚戒指，請你判斷它是否是純金做的。請設計你判斷其密度的步驟。並解釋如何操作，使用的器材等。

（注：本題評分按照錯誤(Incorrect)，部分正確(Partial)，正確(Complete)三個等級來評定。）

[點評]本題是 12 年級考察物理實驗操作的問答題。需要學生自行設計實驗步驟，自己考慮如何使用器材，步驟不限，具有一定的開放性。本題相當於 12 年級 194 分的要求。在 12 年級分數達到 194 分的美國學生中，有超過 65%的學生能正確回答本題。

例題 6：（12 年級）

用工具在地面上觀測太陽，發現 1 月份的太陽會比 7 月份的太陽稍微大些，請從下面選出正確的原因。（ ）

A 地球繞著太陽沿著橢圓軌道運轉，在 1 月份比在 7 月份更靠近太陽。

B 地球的直徑不是常數，在赤道上會大些，並且冬季也是如此。

C 地球的軌道不像其他行星在同一個平面。

D 地球的旋轉軸不是垂直於軌道平面，而是傾斜一個角度。

答案：A

[點評]本題是有關地球科學的 12 年級的一道選擇題。本題屬於 12 年級的優秀等級的題目，相當於 12 年級 223 分的要求。在 12 年級分數達到 223 分的美國學生中，有超過 74%的學生能正確回答本題。

從以上 6 道例題可以看出，「國進評」的測試題並不難，但有一個很重要的特點就是，無論是選擇題還是問答題，都是儘量從學生身邊的問題入手，考查學生在具體的問題情境中解決問題的能力。並且問答題答案並不唯一，往往具有一定的開放性。這就避免學生死記硬背一些知識,這一點顯然很有意義。當然，也有些題目的性質是近似硬背知識，例如：例題 1.

評論及總結

作爲全美唯一從全國不同地區，從三個不同年齡段採集學業成績的資訊體系，「國進評」持續時間長達數十年，其樣本數量遠大於美國其他一些代表性的學生成績調查，如美國

有名的教育縱向調查（NESL）樣本容量僅占「國進評」的 10%-20%。並且「國進評」樣本包含不同種族的學生和學生的家庭、教育背景等資料。它能基本反映美國初等和中等科學教育的現狀，能揭示出美國各州、各地區在教育工作上的不足，爲改進教育工作提供參考，是當之無愧地 “美國國家教育的晴雨錶”。

除此之外，我們認爲最近十年的 3 次「國進評」科學評估，還具有以下幾個特點值得關注。

（1）「國進評」科學評估強調概念理解；從「國進評」的評估框架和測試題可以看出，概念理解在各個年級均占 45%以上，「國進評」科學評估尤其強調那些學生對身邊的科學概念的理解(Resnick, 1987)，突出科學概念在日常生活情景中的運用，而不僅僅是考核學生對概念的識記(苑大勇, 2007)。

（2）「國進評」科學評估強調科學探究的過程；「國進評」科學評價強調概念理解的同時還非常強調科學探究，把科學探究作爲學生獲取知識的工具(朱行建,2007; 趙保鋼,2007)。並且年級越低，越強調科學探究（4 年級最高占 38%，12 年級最低占 28%）。希望學生在一系列探究的過程中，用親身的體驗和行爲來加深對科學概念的理解，並在探究的過程中將科學的概念應用在新的情況中，以便更好地實現知識的遷移。

（3）「國進評」科學評估重視科學方法的掌握；「國進評」的評估框架列出的一些探究性的科學主題中，除了需要學生尋找資料和資料，確定、篩選有用的資訊外，「國進評」科學評估還強調學生的解讀和使用圖表的能力。因爲能否高效地獲取資訊，是學生未來能否很好適應資訊爆炸的社會的重要方面。例如 4 年級的評分標準圖中就多次提到根據圖表資料判斷、解釋問題，問答題和實驗題更是經常需要學生通過圖表有效的表達結果。

（4）「國進評」科學評估強調對科學知識應用技能的真實性評估；「國進評」科學評估考核學生能力比較全面、合理。它一改美國傳統的客觀性試題一統天下的局面，採用了選擇題、問答題和實驗操作題。用選擇題測評學生對重要事實和概念的掌握情況，用問答題重點考察學生理解、分析、應用、傳達科技資訊的能力，採用實驗操作題，真實性地評價學生觀察、動手實驗等能力。既有紙筆測驗，也有實驗動手操作，對難以考察的探究能力、理論運用於實踐的真實水準是一個很好的途徑。

（5）注重考察學生的思維過程，尤其是學生的高層思維能力；「國進評」科學評估中的不少試題具有開放題的性質，評估測試的過程中，不強調答案的唯一性，而是盡可能發散，盡可能讓學生思考，並給其選擇的權利，選擇合適的測量工具，設計自己的實驗步驟。重點查看學生有沒有掌握科學的思維過程，用科學的思維在相同或不同的領域裏處理新的問題。鼓勵學生自己建立、測試和修改理論模型，提倡學生獨立思考，引發學生的高層次思維。

同時，「國進評」試題的設計，按照專案反應理論，採用了大型題庫和矩陣技術，既盡可能全面地考核學生所學，又不致增加學生的測試負擔，使得評估更加高效。

當然，「國進評」科學評估也有其自身的不足：如評估結果發佈的時間間隔太長（18—24 個月），評估標準還可以進一步改進。同時，由於「國進評」本身的設計模式決定了僅

能收集 4、8、12 這三個年級的學校教學實踐的同期資訊，卻無法知道學生在別的學年的學習體驗（例如學生前一個學年已有的學習經驗），使得其比較基點不完全相同，這是它無法克服的弊端。

從美國「國進評」的結果來看，多項社會經濟因素，包括性別、父母教育程度、種族、學校類型、地區、家庭收入等都會與學生科學成績之間存在明顯的相關性。此外，從該三次「國進評」成績的縱向比較，我們也可看到美國科學教育的多個層面都在不斷地改進中。以上的社會經濟因素引致的差異也有不少改善的趨勢。

對於這些結果和當中的原因是否適用於中國內地的科學教育情況，都是值得我們(包括一線科學教師、教育官員和科學教育學者/研究員)參考、反思及/或作出深入的進一步研究。但是，在採納或引用美國「國進評」的結果或評估方法前，我們必須小心謹慎地考慮當中的社會文化的差異、教育制度和課程上的差異，評估目的以及評估結果的不同和可能用途。

綜上所述，「國進評」科學評估的經驗和結果告訴人們：好的教育評估應該既能起到全面瞭解教育現狀，檢驗教學效果的作用，又能充分發揮正面導向作用，引導科學教育：重視學生對科學概念的理解，提倡探究式的科學學習，增加學生親身體驗，從而真正把所學運用於日常生活的問題解決中。讓科學不再是抽象的學術知識，讓學生靈活地掌握新經濟要求的知識和技能，以便學生更好地適應未來社會。

參考文獻

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.

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.

周紅，美國國家教育進展評估 NAEP 體系的產生與發展，外國教育研究，2005 年第 2 期。

苑大勇，美國《2005 年城市地區學生科學能力試驗性評估》報告簡述，教育研究，2007 年 3 月。

朱行建，國際教育評價中的科學探究能力測評簡介及啓示，課程教材教法，2007 年 2 月。

趙保鋼，美國全國教育進步評價（NAEP）中的科學探究，物理教學探討，2005 年第 7 期(上半月)。