

英國 APU 科學成就調查

李揚津

香港教育學院

電郵：ycllee@ied.edu.hk

收稿日期：二零零七年十二月廿一日(於二零零八年六月廿八日再修定)

內容

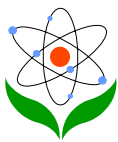
- [摘要](#)
- [背景](#)
- [調查的理論依據和基本框架](#)
- [評量的主要項目](#)
- [擬題原則](#)
- [考題內容](#)
- [調查對象及取樣方法](#)
- [試卷的設計](#)
- [調查結果的分析](#)
- [調查的優點與不足](#)
- [影響及啓示](#)
- [參考書目](#)

摘要

本文探討英國 APU 科學成就調查的成立背景，理論框架，評量項目，測考內容，擬題原則，及測考模式；並就調查結果作扼要分析，包括學生在科學過程技能方面的成就表現，以及學生的表現與各種背景因素的關聯；最後討論該調查的優點及不足之處，並探討其對目下英國國家科學課程的影響。

背景

英國 APU 科學成就調查始於 1975 年，其時英國教育及科學部成立了 APU(Assessment Performance Unit)，主要負責評估學生在三個學習領域的成就，包括語文，數學和科學。目的是「促進評量方法的發展，監察在學兒童的成就，與及識別低學業成就的情



況」。根據此宗旨，APU 有四個主要職能(DES, 1988)：

1. 識別和評鑒現有的評估方法和工具，以達致上述目標
2. 資助開發新的評估工具和技術，同時考慮統計和取樣方法
3. 促進地區教育部門和老師之間的協作，以進行評估
4. 識別在不同學習環境下，學生學業成就出現的顯著差異，包括低成就問題，以及將評估結果公諸於教育部門及學校內負責資源分配的人士。

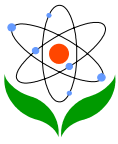
APU 調查背後的基本哲學是以跨學科的觀點，評估學生在不同領域的發展情況，其中包括學生的科學思維發展(DES, 1989)。APU 認為一向以來，科學課程對科學過程技能，包括觀察，推論，提出假說，設計實驗，分析等，未有給予應有的重視。過程技能一般只被視為教授科學概念的手段，而對其內在價值卻未予以肯定(Nuttall, 1992)，這種重概念，而輕技能的觀念亦在課堂評量的設計上得到充份反映。APU 調查的創始就是希望扭轉這種偏差現象，將科學的評量重新定位。APU 成立了不同的學科工作小組，負責各學科評量的監察工作，科學的部份主要是由來自列斯大學(Leeds University)及倫敦大學英皇學院(Chelsea King's College London University)的兩個科學教育專家小組負責。

調查的理論依據和基本框架

APU 工作小組認為科學是一套思考和解決問題的方法，因此，科學的成就評量應強調科學過程技能的表現，而非知識的記取，但該組織亦確信科學過程技能是不可能抽離於科學內容或科學的學習情境而獨立存在的，所以在擬題時，應同時考慮過程技能和科學知識的關係。

APU 採取範疇取樣的方法，即是先訂出要評量的項目或範疇，然後就每一個主項目和子項目建立試題庫。每次調查時，從各個試題庫中隨機抽出相同數量的考題使用。根據 APU 的過程技能評估框架，評估項目被分成六個主項，每個主項又分為兩至三個子項目。評估框架共經歷數次修訂，下表概括了各個主要評估項目，子項目及評估方式(Murphy and Gott, 1984, p5)。

主項目	子項目	評估方式
運用圖表和符號表達	· 從圖表及表格中讀取資料 · 以圖表及表格表示資料	筆試
運用器材及量度儀器	· 運用量度儀器 · 估計物理數量 · 根據指示進行實作活動	小組實作評量
觀察	· 進行觀察和分析觀察所得	小組實作評量
分析和應用	· 分析他人提供的資料 · 應用生物概念 · 物理概念 · 化學概念	筆試
設計探究	· 設計探究的不同部份	筆試
進行探究	· 進行整項探究	個人實作評量



以上的框架主要應用於 13 和 15 兩個年齡組別，至於 11 歲組別，在「運用器材及量度儀器」的主項中，只評量學生運用簡單器材和量度工具的能力；在「應用」方面，此組別並沒有將科學概念區分為生物，物理和化學三方面(Harlen et al 1984)。

以上的評估框架為設計評估工具提供了一個操作基礎，也同時用於報告評估結果，以便於改善教學。正如 APU 工作小組所言(DES 1989)，這套以科學過程技能為主的評估框架，並非建構自任何探究科學本質的哲學觀，亦不是以強調發展階段的學習心理模型為依據。但工作小組深信這些評估項目是科學教育應達致的重要學習成果，其觀點是將科學視為以實驗為本位的探究活動。以下就各主要項目，子項目及相關的考題類型作更詳細說明，考題的例子引自 Murphy and Schofield. (1984) 和 Harlen et al (1984)。

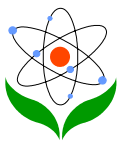
評量的主要項目

第一主項「運用圖表和符號表達」是評量學生運用數學和科學的獨特溝通方式的能力，其中包括兩個子項，一是「從圖表及表格中讀取資料」，例如從折線圖讀取一條河的闊度和深度，及根據食物網、電路圖、化學方程式或切面圖找出相關的資料；二是「以圖表及表格表示資料」，例如將一組在不同時間內所量度得的植物高度的資料以圖表方式表達，以顯示植物的高度如何隨時間而改變(Harlen et al, 1984, p14-16)。

第二主項「運用器材及量度儀」是評量學生對量度方法的認識及應用能力。這個項目的評量是以小組實作方式進行，一組學生會被安排在同一時間內輪流完成指定的任務。這主項包括三個子項。一是「運用量度儀器」，考題包括利用量筒量度液體的體積，又或是一塊細小固體的體積，利用杠桿秤量度對象的品質，及利用安培計量度電流等。第二個子項目是「估計物理數量」，包括估計箱子的體積、葉片的面積、鐵線的長度、包裹的重力、皮球的品質等。第三個子項是評量「學生根據指示進行實作活動」的能力；學生會按照指示進行一系列的實驗操作，包括轉移指定份量的固體化合物和稀酸，將固體化合物和稀酸在試管中混合，然後利用本生燈將試管加熱，最後過濾試管中的液體。學生進行這活動時，老師會在一旁觀察，然後在一張清單上，評量學生進行每個步驟的準確性，並記下學生欠缺哪些方面的能力，以妨礙他完成該項任務。

第三主項是評估學生的觀察能力，及能否辨認出觀察結果所顯示的特別模式，並以這些模式作為預測的基礎。這項目是以實作任務作為評量的依據，例如，其中一項任務是要求考生觀察青蛙的五個成長階段，然後說出每個階段之間有何相同及不同的地方。另一題是要學生將三支載有不知名化合物的試管加熱，然後記錄各種化合物的變化，最後比較各支試管的共通變化(Murphy and Schofield, 1984, p31)。

第四個項目評估學生能否根據所給予的觀察資料描述該等資料的特點，根據資料作出假設，應用已習得的科學概念，及評估有關結論的有效度。這個項目分為兩個子項，第一個子項是分析已集得的資料，包括分辨不同程度的推論，例如，其中一題要求學生根據圖片資料，分析樹木的高度與年輪數目的關係 (Murphy and Schofield, 1984, p22)；另一題展示一幅圖，圖中顯示在暖爐上擺放著一盆枯萎了的植物。學生需要在不同選項中揀選一項最能代表其所觀察到的情況，而在各選項中，只有一項是基於觀



察，而沒有加入主觀的推論。設計這考題的目的是要評量學生能否分辨觀察和推論的分別(Murphy and Schofield, 1984, p23)。

第四個主項中的第二個子項，是評量學生應用科學概念包括物理、化學和生物概念的能力，例如，能否根據圖表所展示的資料，解釋溪流中的含氧量在不同時間內出現變化的原因。另一例子是要求考生解釋為何手電筒出現銹蝕現象後，燈泡便不發亮，但如去除鐵銹，便可操作如常(Murphy and Schofield, 1984, p24-25)。

第五主項是設計探究，這項目的評量主要涉及三種不類型的活動，第一是設計整項探究活動，例如利用盛載了熱水的罐子，比較不同物料的保暖效能(Murphy and Schofield, 1984, p12)。第二是提出可測試的假設，這部份考題的重點是評量學生能否根據某些聲稱，提出一些可以被驗證的假設，這些聲稱包括：羽毛比鉛輕，啡色的蛋比白色的蛋好等等(Murphy and Schofield, 1984, p18)。第三類活動是設計探究活動中的某些步驟，例如提出在某個探究活動所需要控制的變因，又或是評估某個步驟是否存在問題等等。

最後的一個主項目是進行探究，目的是評量學生對各個評估項目的綜合運用能力。這個項目被視為科學教育的一個重要的目標，因為很多證據指出，縱使學生能夠在個別技能包括設計實驗有滿意表現，亦未必能夠將計畫付諸實行(APU, 1979)。這個主項是透過實作方式進行評量。例如，其中一題假設考生流落在一個荒山之上，要冒著強風，寒冷及乾燥的天氣，他需要找出用哪一種物料做的外套最能保暖。題目會給予考生必需的提示，例如可以利用鐵罐和熱水模擬人的身體，及用電風筒吹出的風來模擬強風等(Murphy and Schofield, 1984)。

以上的六個主項目是建基於小組所提出的一個解決科學問題的模型(圖三) (Gott and Murphy, 1987, p24)，這個模型為以上的評量項目提供了一個理論基礎，當中包涵了解決科學問題所涉及的主要探究步驟，及反映出各步驟之間的關係。

擬題原則

考題描述句

在探究技能方面，為了更細緻地分析每一考題所評量的技能，小組根據每一個子項目的要求，設計了一系列相關的「考題描述句」，例如：

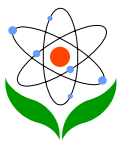
‘從圖表讀取資料’

在一個已提供的資料圖表、典型或非典型的符號中，按指示讀出資料。在一個表示出一個過程或一串有關連的事件或關係的圖表中，按指示讀出資料。

在一個水準或垂直棒形圖中，按指示讀出資料。

在一個圓形圖中，按指示讀出資料。

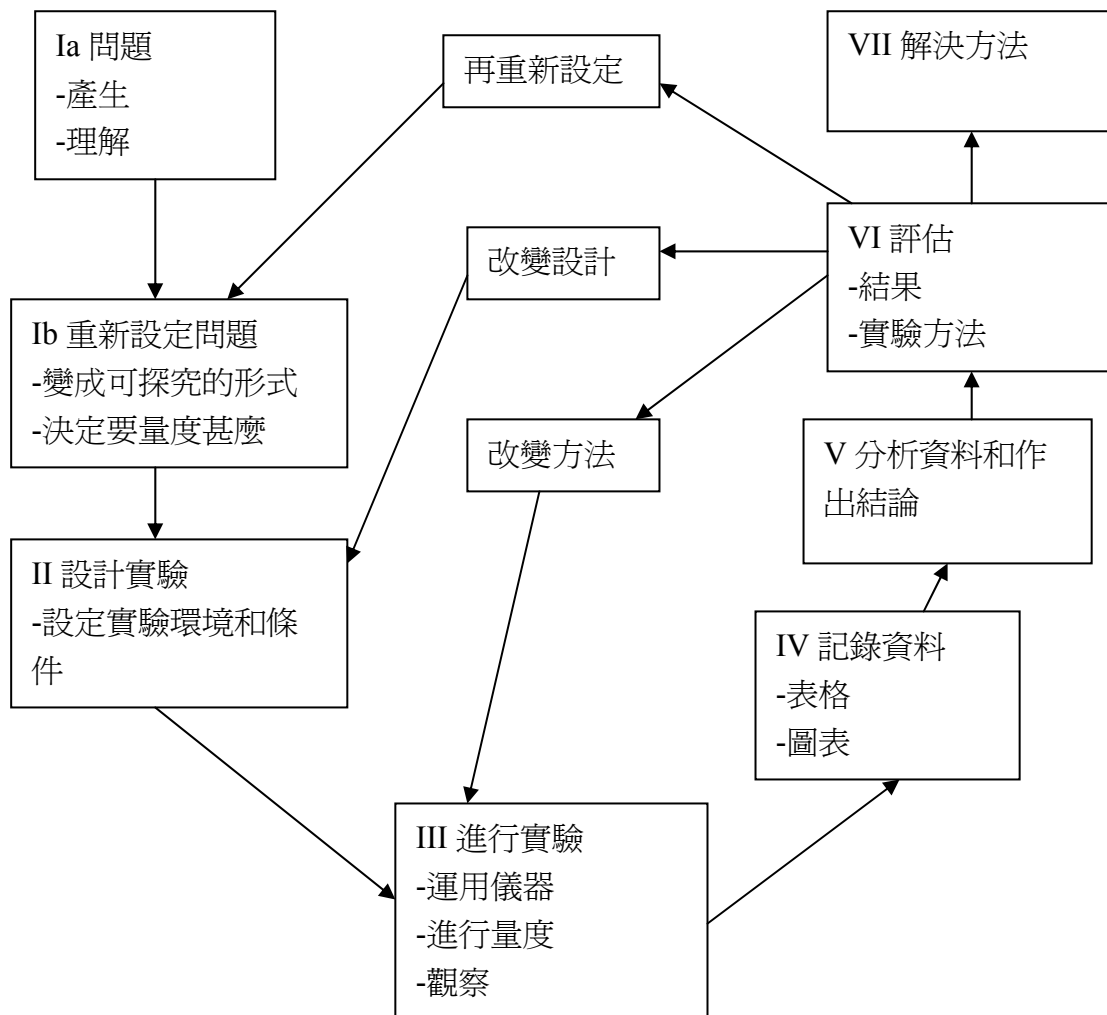
在一個曲線圖中，按指示讀出資料。



在一個圖表及有關的說明中，寫出兩軸的變數。

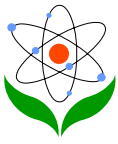
在一個用座標表示的線條或點中，讀出指定的一點座標。

(DES, 1989, p129)



擬題時考慮的因素

爲了更有效地評量學生進行科學探究的能力，在擬定探究問題時，小組考慮了多方面，包括探究活動的目的，性質，內容，情景，對概念的要求，及對科學探究過程的認識等六項要素。首先，評估活動須反映不同的探究目的，例如，有些考題著重比較不同物質的性能，另一些旨在發現變因之間的關係。在活動性質方面，考題分爲寫作或實作兩大類。內容方面，是指活動所提供的資料或器材，包括試題，資料或實驗用品及器材等。情境是指問題是與日常生活有關，還是以純科學爲題材，後者又分爲物理，化學或生物三個範疇。擬題時還須考慮題目是否要求學生對某些概念有基本的掌握，因爲這可能會影響學生在辨別變因，控制變因和使用儀器等方面的表現。最後，小組亦會考慮試題是否要求學生對科學探究的過程有起碼的認識，例如懂得分辨不同



種類的變因，有系統地操縱變因，選擇量度方法，控制變因，從資料中識別重要的變項等等。

工作小組沒有把態度方面的目標列入評估範圍之內，這是因為大眾對態度的理解並不盡同，態度既可指科學態度或科學精神，也可以理解為對科學的態度。前者包含客觀，堅持，批判等；可是，這些態度的特質既不容易清楚界定，亦難以作有效的評估；對科學的態度方面亦存在分歧，對科學的態度究竟是指對某個科學領域的態度，還是對科學於社會的作用的態度，抑或是對科學家的態度等 (DES, 1989)。但基於公眾的要求，小組在 11 歲的組別中加入了一個小規模的調查，以瞭解學生在進行科學探究時所表現的科學態度。而在所有組別中，也利用了調查方式，瞭解學生對學習科學的興趣。

工作小組認同技能的運用很難獨立於學科的內容之外，即是說，學生在一個特定情境中運用技能時，他的表現必然會受到自己對該情境的認識所影響。因此，小組亦同時訂出一系列科學內容範疇，作為擬訂考題的基礎。所有不同年齡組別的考題都是按照這些內容範疇而設計，以便於比較不同年齡學生的進展情況。(DES, 1989)。

考試內容

考題所涵蓋的科學內容範疇

各年齡組別的考題所涵蓋的內容範疇包括以下六個主要部份 (DES, 1989):

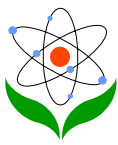
1. 生物于其生存環境的互動
2. 生物與生命過程
3. 力和場
4. 能量轉變
5. 物質的分類和結構
6. 化學作用

每一內容項目又分若干個子項；以第一個項目為例，共分為四個子項：

- A. 生物的互相依存
- B. 物理與化學環境
- C. 生物的分類
- D. 生命現象背後的物理與化學原理

根據不同年齡組別所涵蓋的學習內容，每一子項都訂出了一系列需要評量的概念，那些適用於年紀較小的組別的概念自動被納入較年長的組別之中。以「生物的互相依存」這子項為例，11 歲的組別只包含兩個概念：

- 生物在很多方面互相依存；
- 有些動物以植物為食，有些則以其他動物為食，但所有動物最終都要依賴植物提供食糧」。



在 13 歲組別，增多了以下的一個概念：

- 綠色植物利用來自太陽的能量透過光合作用製造食物。

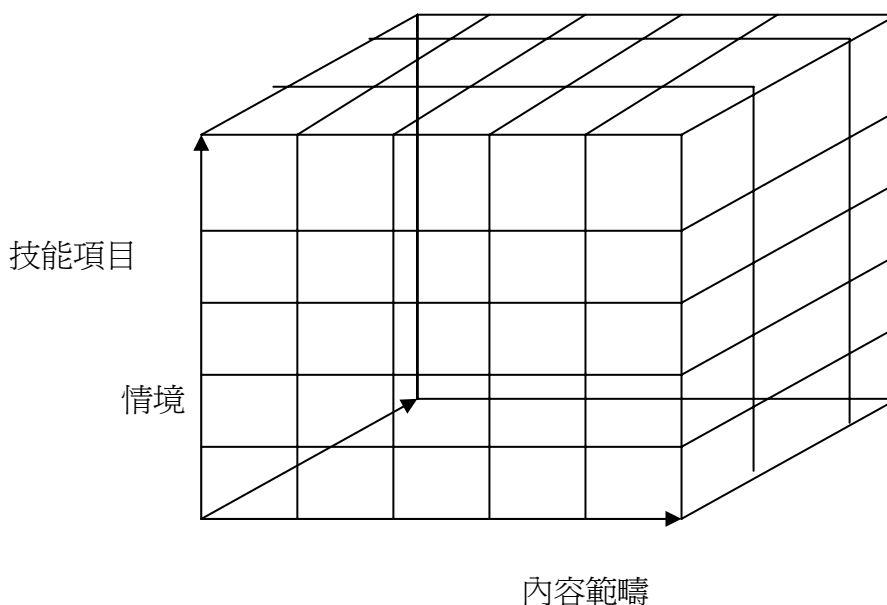
到 15 歲組別，再增添四個概念：· 在一個食物網中，如果任何一部份出現變動，其他部份都可能受影響；

- 食物網出現變化的原因可能是消費者和生產者之間的平衡狀態發生了變化，或是因無機環境發生變化而造成；
- 生物生活於群落之中，而在其所屬群落中，每一生物都佔有一個最適於其生存的位置；
- 競爭和捕食有助維持一個群落中各個種群數量的平衡。

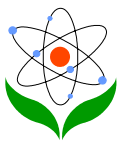
考題的情境

此外，小組亦希望監察學生在整個學校課程中的科學成就，而不單只是科學科上的表現(APU, 1979)。因此在擬題時融入了三類不同的情境：一是與科學課有關的，二是與其他科目有關的，而三是與日常生活有關的。例如，要評估考生是否懂得利用絕緣的方法去保溫，會要求考生研究不同物質的導熱性(與科學課有關的情境)，或指出未發明電冰箱之前，人們是利用甚麼方法令東西保持清涼(與歷史科有關的情境)，又或是找出保持家居溫暖的方法(與日常生活有關的情境) (APU, 1979, p.5)。

基於以上所說，每一考題都包涵了三個向度：過程技能，內容範疇，及問題情境。這三個向度可以組合成一個三維框架(APU, 1979, p.4)，作為制訂考題的基本指引；這框架亦有助於描述每一考題在整個評量中所發揮到的作用。



雖然小組有意透過是次評估，深入瞭解學生在橫跨不同概念範疇及情境下的技能成就表現，但如果按照這個三維框架的每一個方格的要求擬題，題目的種類便非常龐大，



處理上亦異常複雜。因此工作小組決定放棄以情境來將考題分類(DES, 1989)，但仍在考題中滲入不同情境，以保證評估的有效度。小組亦決定除了第四個有關應用相關科學知識的技能項目外，其他考題不應以學生對任何一項內容範疇的認識作為答題的必需條件。

調查對象及取樣方法

根據 DES(1989)，是次評估的對象是 11，13 和 15 歲就讀於英格蘭，威爾斯和北愛爾蘭的學生。第一次大型取樣調查始於 1980 年，以後每年進行一次，直至 1984 年完成最後的一次。調查的目的是監察每組的平均表現，及這表現是否會隨時間而改變。在每個調查年度，每個年齡組別都有約來自 500 至 1,000 間學校的 12,000 至 16,000 名學童參與。在首年的調查，每所學校派出 11 歲學生 8 名和 13 和 15 歲學生各 9 名參予評估。抽樣的方法是挑選每月首日出生的學生為評估對象。在其後的調查，每個年齡組別的參與學生的數目都有所不同。由於學生和考題都是以隨機方式抽樣，因此兩方面的樣本都應具有代表性。

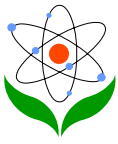
試卷的設計

測考模式

由於評估所牽涉的內容項目眾多，每次測試只能評估部份項目。每個子項目都設有一個試題庫，部份試題適用於兩個或三個年齡組別，以便於比較跨年齡的成就差異。每個子項目的試題庫有約二百條考題，而每次評估測試只使用 45 至 60 題不等，然後再以隨機抽樣方式，抽取約 15 至 20 題印成試卷予學生作答。每次調查約使用 30 份不同的試卷。11 歲組別的評估時間約 45 分鐘，而其他組別的時間為一小時。在筆試前或後，會抽取部份考生進行實作測試。

大部份考題是以筆試方式進行，部份項目例如使用量度儀器，及進行觀察，則採用實作測試方式。在實作測試，每題所需的儀器會被放置於考場的一個區域內，考生輪流到不同場區作答(Schofield et al, 1982)。在評估進行探究這主項時，除了根據考生的書面答案外，亦會參考觀察員對考生的反應所作的記錄。以書面作答的題目類型包括多項選擇題，简答题及開放性的問題。多項選擇題的好處是可以縮短作答的時間及利用機器評分，也可以避免因評卷員的錯誤判斷而出現誤差，但缺點是可以讓考生猜測答案。

實作測試有兩種執行模式，第一種是迴圈方式，各考生在考場內輪流進行多項實作活動，包括運用儀器量度及觀察，器材由大會向考生提供，雖然是實作測試，但考生須以書面方式作答。第二種是個別測試，適用於評量第六主項，即進行整項探究。個別考生須單獨接受評估員測試，評估員先向考生提出問題，引導考生進行實作探究，評估員會在旁觀察並記錄考生所用的方法，事後評估員需要填寫一份評估清單，以記錄考生在探究中的表現，這種模式主要應用於進行探究這項目上，有關評估實作測試的詳細程式可參考 Welford et al (1985)，Murphy and Schofield (1984)，Murphy and Gott (1984) 及 Gott and Murphy (1987)。



每一考題的分數由零至三分，而多項選擇題則為零至一分，由於選擇題的分數較為劃一，可以容許考評當局在每次評估時，採取較靈活的分層任意選題方式準備考卷。這樣，每份考卷的試題組合便不相同，可以避免學校刻意操練學生應試，令評估更為公平。

學校問卷

此研究亦設有學校問卷，目的是收集各參與學校向學生提供的科學教育的資料，以便於找出學校所提供的科學教育與學生的成就表現之間的關聯。問卷的內容包括科學教學的師資和資源，進行科學教學的目的，對科學教育的重視程度，分配予科學教學部門的財政資源等。11 歲和其他年齡組別問卷內容有很大差別，這是因為當時英國的小學所推行的科學教學存在頗大差異，但到了中學階段，科學課程漸趨統一。以下是用於不同年齡組別的學校問卷的一些題目例子(Driver et al, 1982; Schofield et al, 1982)。

11 歲組別：

- 校內是由哪些人員負責科學教學？
- 科學課是否包含於校內學習活動之內？
- 進行科學教育是否學校的方針？
- 科學是時間表上編定的科目，還是占主題教學中的一個部份？
- 10/11 歲的學生用於學習科學的時間占全部課堂時間的百分之幾？
- 科學活動有甚麼目標？
- 學校對以下科學活動的重視程度。

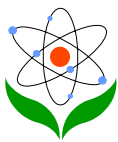
例如：

- 小心地跟隨著咭紙、書本或黑板上的指示去做
- 在課堂上寫下觀察紀錄
- 製造一個良好的書面記錄
- 決定一個他們想調查的問題
- 在真正測試之前估計測量讀數
- 小心進行直接的觀察
- 設計自己的實驗

(問題經簡化)

13 和 15 歲組別：

- 校內有多少名老師擁有正規的科學資歷？
- 校內共有多少間實驗室？
- 學校的科學部共有多少名實驗室技術員？實驗室技術員的工作時數有多少？
- 老師有沒有使用校外資源，例如參觀工業機構，到野外學習中心進行活動等？



- 在一個迴圈周內，共有多少堂科學課？
- 在學校的經費中，科學部所獲得的分配有多少？

學生問卷

學生問卷的設置主要是為幫助分析學生的成就表現，以貫徹 APU 的第四項職能 — 識別學習環境與學生學業成就差異的關聯。問卷旨在瞭解學生多方面的背景資料，例如：性別，種族，修讀科目，職業取向，對有關科學議題的興趣；問卷亦問及學生課餘的興趣及嗜好，這是基於越來越多研究發現，學生的課餘活動與他們學習科學的興趣有關。

效度和信度

為提高評估的效度和信度，小組邀請了科學教育的專家 包括大學教授，教師，考試局成員，督學等審議各範疇的主要項目，子項目和考題描述句，並試行將考題按照小組所設定的框架分類，從而評估框架的有效度，以及各主項目，子項目和考題描述句的清晰度，如有需要便提出修改，透過專家的共識，令評估框架更臻完善。

小組亦進行了先導試驗，以分析考生在不同子項目上的表現的相關性。每一考題草擬後必需符合兩個條件才會被納入試題庫之中，一是專家小組一致評定為具足夠效度；二是該考題的性質必需與其中一個考題描述句相符。此外，評分準則亦經標準化的過程。每位評卷員都是經過特別培訓，以確保評分過程的可靠性。但如要保證評量實踐性考題的客觀性及信度時，則較為困難。這方面惟有依賴先導試驗，以減少由不同評估員作評量而出現的差異 (APU, 1979)。

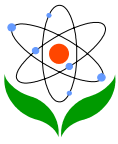
對於進行整項探究的評量分析，小組針對其所設的探究框架下的每一個步驟擬出評量的仔細標準，然後統計達至這些標準的學生人數的百分比，以瞭解學生在不同步驟中的成就表現(Gott and Murphy, 1987)。

調查結果的分析

本調查結果分析可分為兩個部份。第一部份是報告和分析學生的成就表現；第二部份分析學生的成就表現與各種背景因素的關聯。

學生的成就表現

小組除了報告和分析學生在個別考題的表現外，還對學生在個別子項目，項目，及整體調查的表現作出概括性的分析；對個別考題的表現，報告列出了學生所達到的能力或分數的百分比；就每一子項目及項目，報告總結出各年齡學生的成就表現。以 13 歲組別於 1980-81 的調查為例，在利用圖表及符號表達方式這項目之下的其中一個子項 - 利用圖表及表格表達資料，很多學生對設定兩軸的刻度感到困難。就這項目的整體表現而言，學生能夠運用不同的資料表達方式，但如果學生同時要面對多種資料變化時，表現便較差；學生亦不大能將三維圖像轉譯成正統的二維圖形，例如線路圖和切



面圖。這個年齡組別的整體調查果亦顯示，學生在運用圖表及表格，和應用生物概念這兩個項目的成就存在很大差異，前者遠較後者為優 (DES, 1984)。

學生的成就表現與各種背景因素的關聯

學校問卷和學生問卷所提供的資料可以用來分析影響學生成就的因素，以便於解釋評量所發現的趨勢。根據 DES (1989)的報告，這些因素包括學生年齡、性別、所修讀的學科、班級人數、學生在校內接受教育的情況，以至學校對科學課程的安排，投放的資源等等。針對學生的因素而言，調查結果顯示，在相同考題中，學生成就是隨年齡而遞升；學生的表現亦存在性別差異的情況。以應用物理學概念這子項為例，在三個年齡組別中，男生的表現都較女生為優；不過，在分析觀察結果方面，卻出現相反的情形。此外，調查亦顯示男女差異的情況是與學校所屬的地區有關，例如在觀察和分析觀察結果方面，威爾斯的 15 歲組別的男學生比女學生有較佳表現，來自英格蘭和北愛爾蘭的男生卻剛好相反。綜合各個項目的整體表現，北愛爾蘭女學生的表現明顯地比其他地區的女學生優勝。

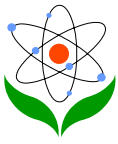
針對學校因素而言，學生于子項中的表現與同類活動在學校的普及程度呈正相關。但有些結果卻是出乎意料之外，例如，調查找不到證據顯示，學生學習科學的時間越長，成就便越高；事實上，較平庸的學生學習科學的時間越長，應用概念的成就越低。此外，雖然位於近郊的學校所獲分配的資源較少，但是這類學校的學生的表現卻比城鎮學校的學生為好。這些結果足以顯示出學生的成就未必與學校因素有關。另一個出人意表的例子是在 13 歲學生中，班中的人數越多，學生的成就越高，這可能是因為學校早已將能力較差的學生編入小班之中，這個解釋亦與學校的情況吻合。

以上例子說明了即使學生的成就與某種背景因素成正相關，也不能證實它們之間存在必然的因果關係。就正如調查發現修讀物理科的學生比念生物科的表現較佳，並不一定表示這是修讀不同科目的後果，而可能是因為成就較高的學生較傾向於揀選物理科而非生物科，當然小組亦不排除學生的表現可能是受到學科的不同教學方法所影響 (DES, 1989)。

出乎意料之外，學生在實作探究的表現比較他們在應用科學概念方面更為優勝，縱使課程偏重後者而忽略前者。根據小組的分析，學生在實作探究的表現可能是受五個因素所影響(Gott and Murphy, 1987)。第一是此類問題並不要求學生以文字作答，因此學生的表現不受語文障礙的影響。第二，此類實作探究提供了不少有關如何探究問題的提示，令學生不會感到問題難於理解，而缺乏此類提示正是文字題令學生感到困難的原因。第三，由於學生有機會不斷嘗試，他們可以根據結果而修訂實驗方法。第四，這些實作測試很少要求學生運用科學概念。第五，大部份實作測試是與日常生活有關，因而鼓勵學生作出嘗試。

調查的優點與不足

APU 可以說是開創了英國以技能為本的評量調查的先河，調查有系統地將科學探究對技能的要求有系統及具體地分為若干主項目和子項目，並以這些項目作為命題的基礎，以保證考題的效度，成為日後此類調查的濫觴。調查亦將學生的成就表現與學生



及學校的背景資料作比較，以辨析其中的關係所在，從而辨析一些可能影響學生成就的因素。

但是，本調查亦有很多不足之處，有些是與調查的目標和可行性有關，是關乎理論與實踐的問題；有些是與問題或問卷的設計有關；亦有些是與其調查方法有關，關乎信度的問題。首先，APU 調查主要是希望監察在學兒童的成就，與及識別低學業成就的情況，但調查的分析結果只能反映學生在本調查各項目和子項中的表現，而不能提供學生成就的標準或指標，因此很難對學童的整體成就作出判斷。

第二，正如 DES(1989)指出，調查的評量框架主要反映參與調查計畫的專家的共識，背後並沒有很強的理論基礎予以支持，所以難免會存在一些結構性的問題，例如一些子項之間互有重複的地方，包括“運用圖表及表格”與“分析資料，設計實驗與進行實驗”等，因此，學生在不同項目中的表現，並非完全獨立。在擬訂考題的過程中，子項目仍被不斷修改，以避免重複，令人對這些評估結果的效度產生懷疑。

第三，小組亦承認過程技能的發揮，是很受探究的情境和學生對相關概念的掌握程度所影響，事實上，很多考題並不能將內容或概念完全抽離，例如在很多評估觀察能力的考題中，只有那些已具備相關概念的考生，才能取得被評卷者視為正確的觀察結果(Driver et al 1982)。此外，很多子項目也要建基於數學及統計學概念例如比例，機率，平均值等(DES, 1989)。

正因如此，縱使學生在某些考題中有出色的表現，亦不一定表示他們能夠將此等技能遷移至其他問題情境之中，好像在「進行探究」這項目中，學生在不同活動的表現存在顯著差異，可見學生的成就表現是受其概念所影響。

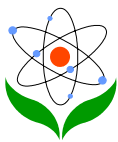
第四，調查嘗試找出與學生成就相關的因素，但從學校和學生問卷所搜集到的資料卻未見深入，調查所得的相關因素研究結果與預期的有明顯分歧，仍需要利用更深入的方法去瞭解不同相關因素之間的互動關係。

第五，由於每次調查所抽取的題目都稍有不同，因此隸屬於相同子項目的考題的分數必須劃一，否則該子項的總分便會出現不一致的情況。為了遷就分數的一致性，學生在個別考題所獲得的分數，就未必能夠完全反映他們在這些考題中的表現。

第六，此成就調查計畫歷時數載，考題也經歷不少變化，例如某些題目原先只出現於一個組別，後來被應用於其他組別，由於考題已經歷變化，所以很難比較在不同年份進行的調查的結果。此外，由於要符合不同組別的要求，同樣的考題在不同組別的評分標準並非一致(DES, 1989)，因此如以調查結果來比較不同年齡的學生的成就或有欠公平。另外，一個更根本的問題是課程，教學方法甚至整個社會都不斷改變，所以有需要按時更新考題，但這卻與公平比較跨年級的進展背道而馳(Gipps, 1988)。

第七，即使各年齡組別都使用相同的評估項目，但此方式有別於追蹤研究，所以並不能用來瞭解學生的真正發展。

第八，正如 DES(1989)指出，任何大型測試都難免出現誤差，所謂誤差是指調查結果與真實情況之間的分別。調查的誤差可以與取樣有關或與取樣無關。根據小組的分析，



與取樣有關的誤差主要來自學生之間，學校之間，及問題之間的差異，前者包括能力上的差異，課程的影響，語文能力，對考題相關情境的熟悉程度等；而學校的差異則包括學校的課程，教學方法，學習氛圍等方面的差異；考題的差異主要是來自考題的不同難度，研究結果的誤差主要受樣本的數量影響，一般來說，樣本越大，出現的誤差越小。

最後，雖然 APU 由始至終強調的是學生的過程技能，但其對學生其他方面的成就表現如概念的理解，及科學態度的忽視卻成爲了外界對它批評的原因，尤其是當時的課程甚少強調過程技能。

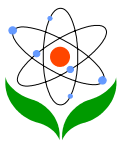
影響及啓示

在英國 APU 調查開創了評量科學過程技能的先河，它的影響可以分爲短期與長期兩方面。短期來說，它加深了教師對科學成就評量的瞭解，及提供了很多在課堂上評量學生成就的意念和素材(Nuttall 1992)。雖然調查未能確實地辨析出影響學生成就的相關因素，但是，它至少反映出一些課程或學習的潛在問題，例如，學生應用科學概念的能力偏低，值得作進一步研究。長遠來說，APU 將科學過程技能的評量具體化及系統化，影響了日後對科學成就的評量，它衍生出 GCSE (General Certificate of Secondary Education)的實作評量部份，及後期用於評量國家課程(National Curriculum)的標準評量任務。由於 APU 只提出了一個評量探究技能的框架，調查結果只能反映不同項目或子項之間的相對表現，所以小組並沒有將各項目的表現轉譯成一個總分數，並以此標示每名學生或整個組別的成就表現(DES 1989)。因此，調查結果未能顯示出學生是否已達到一定的標準，亦未能指出這絕對或相對標準應如何界定。不過，調查結果的公佈開始令外界關注到應否爲不同年齡的學生厘訂成就標準。因爲如果沒有訂出統一的標準，課程發展者便很難厘定清晰的目標，以配合不同年齡組別的發展。基於此，在九零年代初英國發展了國家科學課程，其中一個嶄新目標就是發展學生的科學探究能力，課程石破天驚地提出了一套配合不同發展階段的目標階梯，作爲學校發展課程和評量學生成就的準則，這套準則的誕生與 APU 的調查不無關係，更可以理解爲 APU 對科學探究所持有的理念的延伸 (Donnelly and Jenkins, 2001)。該目標階梯經多年的發展已趨於成熟，雖然還有未盡善的地方，亦未必能夠全面及完全準確地反映學生在探究方面的能力發展，但它對於科學課程已起著兩方面頗爲深遠的影響。其一是學校及教師需要以該目標階梯爲準則，設計針對不同年級或不同能力水準的課程，希望能夠做到因材施教，從而逐步提升學生的水準。其二是讓學校及教育當局以此作爲評量各校各級學生的成就表現的劃一準則。爲配合國家課程的評鑒，英國評估當局發展了一套名爲標準評量任務(Standard assessment tasks, SAT)的實作評量工具，以測試學生的探究技能，雖然後期基於教師工作量及時間上的限制，這套工具很大程度上被書面測試所取代，但毫無疑問，這套評量工具的影響已反映在課堂教學之上。由此可見，APU 所奠下的評估框架和理念還是有一定影響的。這方面的影響並不限於英國本土，還擴展及世界的其他地方。

參考書目

APU (1979). Science Progress Report 1977-78, London: HMSO.

DES (1984) APU Science report for teachers: 3, Science at Age 13. London: HMSO.



- DES (1988). Science at Age 15 – A review of APU survey findings, 1980-84. London: HMSO.
- DES (1989). National Assessment: The APU Science Approach. London: HMSO
- Driver, R, Goot, R, Johnson, S, Worsley, C, and Wylie, F (1982). Science in Schools, Age 15: Report No. 1. London: HMSO.
- Donnelly, J. F., and Jenkins, E. W. (2001). Science education: policy, professionalism and change. London: Paul Chapman.
- Gipps, C. (1988). The debate over standards and the uses of testing. *British Journal of Educational Studies*, XXXVI(1), 21-37.
- Gott, R, and Murphy, P. (1987). Science report for teachers: 9 – Assessing investigations at ages 13 and 15. Letchworth, Hertfordshire: Garden City Press
- Harlen, W., Palacio, D., and Russell, T. (1984). Science report for teachers: 4 – Science assessment framework age 11. Letchworth, Hertfordshire: Garden City Press
- Murphy, P. and Gott, R. (1984). Science report for teachers: 2 – Science assessment framework age 13&15. Letchworth, Hertfordshire: Garden City Press
- Murphy, P. and Schofield, B. (1984). Science report for teachers: 3 – Science at age Letchworth, Hertfordshire: Garden City Press.
- Nuttall, D. L. (1992). Performance assessment: The message from England. *Educational Leadership*, May, 54-57.
- Schofield, B, Murphy, P, Johnson, S., and Black, P. (1982). Science in Schools Age 13: Report No. 1. London: HMSO.
- Welford, G., Harlen, W., and Schofield, B. (1985). Science report for teachers: 6 – Practical testing at ages 11, 13&15. Letchworth, Hertfordshire: Garden City Press.