英国 APU 科学成就调查

李扬津

香港教育学院

电邮：yclee@ied.edu.hk

收稿日期：二零零七年十二月廿一日(于二零零八年六月廿八日再修定)

内容

摘要

本文探讨英国 APU 科学成就调查的成立背景，理论框架，评量项目，测考内容，拟题原则，及测考模式；并就调查结果作扼要分析，包括学生在科学过程技能方面的成就表现，以及学生的表现与各种背景因素的关联；最后讨论该调查的优点及不足之处，并探讨其对目下英国国家科学课程的影响。

背景

英国 APU 科学成就调查始于 1975 年，其时英国教育及科学部成立了 APU(Assessment Performance Unit)，主要负责评估学生在三个学习领域的成就，包括语文，数学和科学。目的是「促进评量方法的发展，监察在学儿童的成就，与及识别低学业成就的情况」。根据此宗旨，APU 有四个主要职能(DES, 1988)：

1. 识别和评鉴现有的评估方法和工具，以达致上述目标
2. 资助开发新的评估工具和技术，同时考虑统计和取样方法
3. 促进地区教育部门和老师之间的协作，以进行评估
4. 识别在不同学习环境下，学生学业成就出现的显著差异，包括低成就问题，以及将评估结果公诸于教育部门及学校内负责资源分配的人士。

APU 调查背后的基本哲学是以跨学科的观点，评估学生在不同领域的发展情况，其中包括学生的科学思维发展(DES, 1989)。 APU 认为一向以来，科学课程对科学过程技能，包括观察，推论，提出假说，设计实验，分析等，未有给予应有的重视。过程技能一般只被视为教授科学概念的手段，而对其内在价值却未予以肯定(Nuttall, 1992)，这种重概念，而轻技能的观念亦在课堂评量的设计上得到充份反映。APU 调查的创始就是希望扭转这种偏差现象，将科学的评量重新定位。APU 成立了不同的学科工作小组，负责各学科评量的监察工作，科学的部份主要是由来自列斯大学(Leeds University)及伦敦大学英皇学院(Chelsea King's College London University)的两个科学教育专家小组负责。

调查的理论依据和基本框架

APU 工作小组认为科学是一套思考和解决问题的方法，因此，科学的成就评量应强调科学过程技能的表现，而非知识的记取，但该组织亦确信科学过程技能是不可能抽离于科学内容或科学的学习情境而独立存在的，所以在拟题时，应同时考虑过程技能和科学知识的关系。

APU 采取范畴取样的方法，即是先订出要评量的项目或范畴，然后就每一个主项目和子项目建立试题库。每次调查时，从各个试题库中随机抽出相同数量的考题使用。根据 APU 的过程技能评估框架，评估项目被分成六个主项，每个主项又分为两至三个子项目。评估框架共经历数次修订，下表概括了各个主要评估项目，子项目及评估方式(Murphy and Gott, 1984, p5)。

主项目	子项目	评估方式
运用图表和符号表达	· 从图表及表格中读取数据 · 以图表及表格表示数据	笔试
运用器材及量度仪器	· 运用量度仪器 · 估计物理数量 · 根据指示进行实作活动	小组实作评量
观察	· 进行观察和分析观察所得	小组实作评量
分析和应用	· 分析他人提供的数据 · 应用生物概念 物理概念 化学概念	笔试
设计探究	· 设计探究的不同部份	笔试
进行探究	· 进行整项探究	个人实作评量

以上的框架主要应用于 13 和 15 两个年龄组别，至于 11 岁组别，在「运用器材及量度仪器」的主项中，只评量学生运用简单器材和量度工具的能力；在「应用」方面，此组别并没有将科学概念区分为生物，物理和化学三方面(Harlen et al 1984)。

以上的评估框架为设计评估工具提供了一个操作基础，也同时用于报告评估结果，以便于改善教学。正如 APU 工作小组所言(DES 1989)，这套以科学过程技能为主的评估框架，并非建构自任何探究科学本质的哲学观，亦不是以强调发展阶段的学习心理模型为依据。但工作小组深信这些评估项目是科学教育应达致的重要学习成果，其观点是将科学视为以实验为本位的探究活动。以下就各主要项目，子项目及相关的考题类型作更详细说明，考题的例子引自 Murphy and Schofield. (1984) 和 Harlen et al (1984)。

评量的主要项目

第一主项「运用图表和符号表达」是评量学生运用数学和科学的独特沟通方式的能力，其中包括两个子项，一是「从图表及表格中读取数据」，例如从折线图读取一条河的阔度和深度，及根据食物网、电路图、化学方程式或切面图找出相关的数据；二是「以图表及表格表示数据」，例如将一组在不同时间内所量度得的植物高度的数据以图表方式表达，以显示植物的高度如何随时间而改变(Harlen et al, 1984, p14-16)。

第二主项「运用器材及量度仪」是评量学生对量度方法的认识及应用能力。这个项目的评量是以小组实作方式进行，一组学生会被安排在同一时间内轮流完成指定的任务。这主项包括三个子项。一是「运用量度仪器」，考题包括利用量筒量度液体的体积，又或是一块细小固体的体积，利用杠杆秤量度对象的质量，及利用安培计量度电流等。第二个子项目是「估计物理数量」，包括估计箱子的体积、叶片的面积、铁线的长度、包裹的重力、皮球的质量等。第三个子项是评量「学生根据指示进行实作活动」的能力；学生会按照指示进行一系列的实验操作，包括转移指定份量的固体化合物和稀酸，将固体化合物和稀酸在试管中混合，然后利用本生灯将试管加热，最后过滤试管中的液体。学生进行这活动时，老师会在一旁观察，然后在一张清单上，评量学生进行每个步骤的准确性，并记下学生欠缺哪些方面的能力，以妨碍他完成该项任务。

第三主项是评估学生的观察能力，及能否辨认出观察结果所显示的特别模式，并以这些模式作为预测的基础。这项目是以实作任务作为评量的依据，例如，其中一项任务是要求考生观察青蛙的五个成长阶段，然后说出每个阶段之间有何相同及不同的地方。另一题是要学生将三支载有不知名化合物的试管加热，然后记录各种化合物的变化，最后比较各支试管的共通变化(Murphy and Schofield, 1984, p31)。

第四个项目评估学生能否根据所给予的观察数据描述该等数据的特点，根据数据作出假设，应用已习得的科学概念，及评估有关结论的有效度。这个项目分为两个子项，第一个子项是分析已集得的数据，包括分辨不同程度的推论，例如，其中一题要求学生根据图片数据，分析树木的高度与年轮数目的关系 (Murphy and Schofield, 1984, p22)；另一题展示一幅图，图中显示在暖炉上摆放着一盆枯萎了的植物。学生需要在不同选项中拣选一项最能代表其所观察到的情况，而在各选项中，只有一项是基于观察，而没有加入主观的推论。设计这考题的目的是要评量学生能否分辨观察和推论的分别(Murphy and Schofield, 1984, p23)。

第四个主项中的第二个子项，是评量学生应用科学概念包括物理、化学和生物概念的能力，例如，能否根据图表所展示的数据，解释溪流中的含氧量在不同时间内出现变化的原因。另一例子是要求考生解释为何手电筒出现锈蚀现象后，灯泡便不发亮，但如去除铁锈，便可操作如常(Murphy and Schofield, 1984, p24-25)。

第五主项是设计探究，这项目的评量主要涉及三种不类型的活动，第一是设计整项探究活动，例如利用盛载了热水的罐子，比较不同物料的保暖效能(Murphy and Schofield, 1984, p12)。第二是提出可测试的假设，这部份考题的重点是评量学生能否根据某些声称，提出一些可以被验证的假设，这些声称包括：羽毛比铅轻，啡色的蛋比白色的蛋好等等(Murphy and Schofield, 1984, p18)。第三类活动是设计探究活动中的某些步骤，例如提出在某个探究活动所需要控制的变因，又或是评估某个步骤是否存在问题等等。

最后的一个主项目是进行探究，目的是评量学生对各个评估项目的综合运用能力。这个项目被视为科学教育的一个重要的目标，因为很多证据指出，纵使学生能够在个别技能包括设计实验有满意表现，亦未必能够将计划付诸实行(APU, 1979)。这个主项是透过实作方式进行评量。例如，其中一题假设考生流落在一个荒山之上，要冒着强风，寒冷及干燥的天气，他需要找出用哪一种物料做的外套最能保暖。题目会给予考生必需的提示，例如可以利用铁罐和热水模拟人的身体，及用电风筒吹出的风来模拟强风等(Murphy and Schofield, 1984)。

以上的六个主项目是建基于小组所提出的一个解决科学问题的模型(图三) (Gott and Murphy, 1987, p24)，这个模型为以上的评量项目提供了一个理论基础，当中包涵了解决科学问题所涉及的主要探究步骤，及反映出各步骤之间的关系。

拟题原则

考题描述句

在探究技能方面，为了更细致地分析每一考题所评量的技能，小组根据每一个子项目的要求，设计了一系列相关的「考题描述句」，例如：

‘从图表读取数据’

在一个已提供的数据图表、典型或非典型的符号中，按指示读出资料。在一个表示出一个过程或一串有关连的事件或关系的图表中，按指示读出数据。

在一个水平或垂直棒形图中，按指示读出资料。

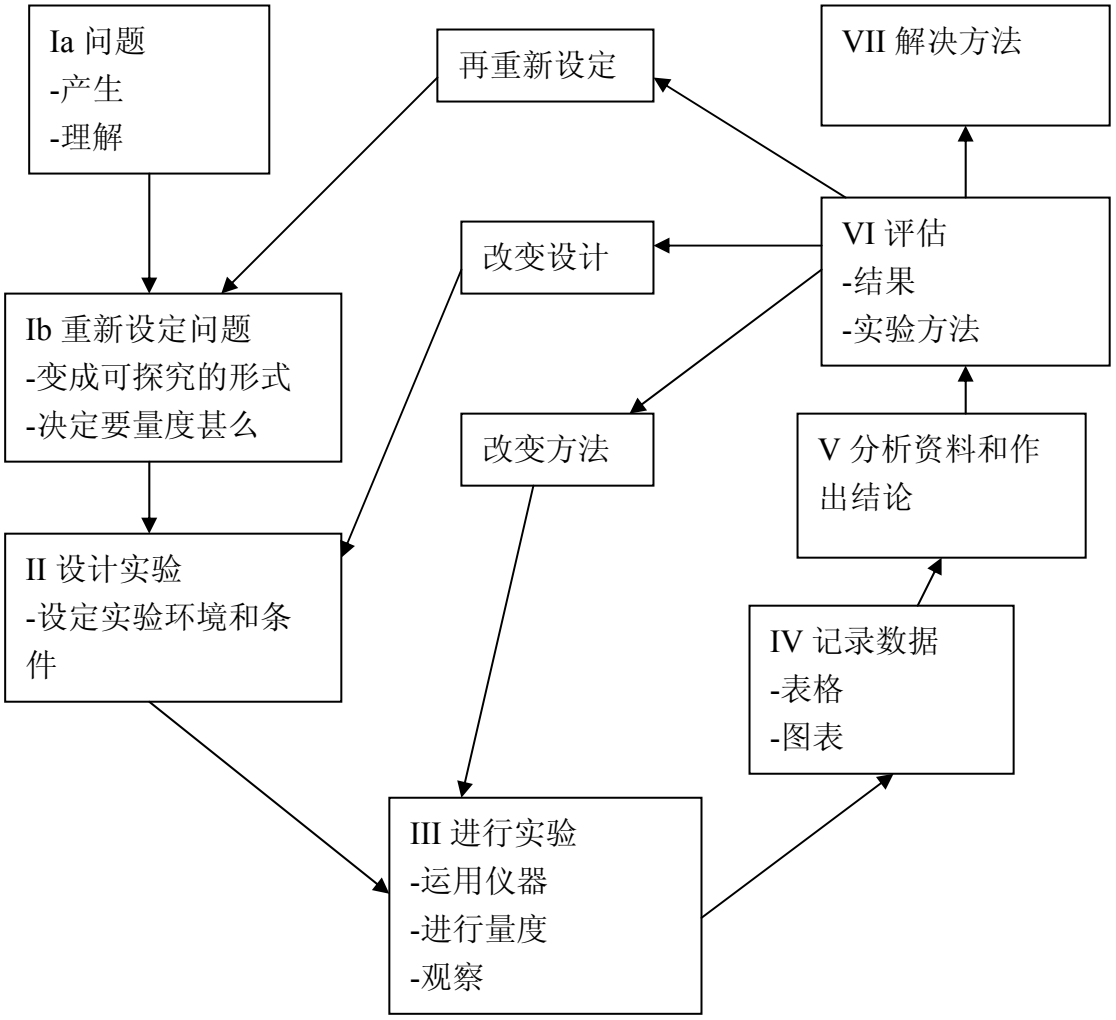
在一个饼图中，按指示读出资料。

在一个曲线图中，按指示读出资料。

在一个图表及有关的说明中，写出两轴的变量。

在一个用坐标表示的线条或点中，读出指定的一点坐标。

(DES, 1989, p129)



拟题时考虑的因素

为了更有效地评量学生进行科学探究的能力，在拟定探究问题时，小组考虑了多方面，包括探究活动的目的，性质，内容，情景，对概念的要求，及对科学探究过程的认识等六项要素。首先，评估活动须反映不同的探究目的，例如，有些考题着重比较不同物质的性能，另一些旨在发现变因之间的关系。在活动性质方面，考题分为写作或实作两大类。内容方面，是指活动所提供的数据或器材，包括试题，数据或实验用品及器材等。情境是指问题是与日常生活有关，还是以纯科学为题材，后者又分为物理，化学或生物三个范畴。拟题时还须考虑题目是否要求学生对某些概念有基本的掌握，因为这可能会影响学生在辨别变因，控制变因和使用仪器等方面的表现。最后，小组亦会考虑试题是否要求学生对科学探究的过程有起码的认识，例如懂得分辨不同种类的变因，有系统地操纵变因，选择量度方法，控制变因，从数据中识别重要的变项等等。

工作小组没有把态度方面的目标列入评估范围之内，这是因为大众对态度的理解并不尽同，态度既可指科学态度或科学精神，也可以理解为对科学的态度。前者包含客观，坚持，批判等；可是，这些态度的特质既不容易清楚界定，亦难以作有效的评估；对科学

的态度方面亦存在分歧，对科学的态度究竟是指对某个科学领域的态度，还是对科学于社会的作用的态度，抑或是对科学家的态度等 (DES, 1989)。但基于公众的要求，小组在 11 岁的组别中加入了一个小规模的调查，以了解学生在进行科学探究时所表现的科学态度。而在所有组别中，也利用了调查方式，了解学生对学习科学的兴趣。

工作小组认同技能的运用很难独立于学科的内容之外，即是说，学生在一个特定情境中运用技能时，他的表现必然会受到自己对该情境的认识所影响。因此，小组亦同时订出一系列科学内容范畴，作为拟订考题的基础。所有不同年龄组别的考题都是按照这些内容范畴而设计，以便于比较不同年龄学生的进展情况。(DES, 1989)。

考试内容

考题所涵盖的科学内容范畴

各年龄组别的考题所涵盖的内容范畴包括以下六个主要部份 (DES, 1989)：

1. 生物于其生存环境的互动
2. 生物与生命过程
3. 力和场
4. 能量转变
5. 物质的分类和结构
6. 化学作用

每一内容项目又分若干个子项；以第一个项目为例，共分为四个子项：

A. 生物的互相依存
B. 物理与化学环境
C. 生物的分类
D. 生命现象背后的物理与化学原理

根据不同年龄组别所涵盖的学习内容，每一子项都订出了一系列需要评量的概念，那些适用于年纪较小的组别的概念自动被纳入较年长的组别之中。以「生物的互相依存」这子项为例，11 岁的组别只包含两个概念：

- 生物在很多方面互相依存；
- 有些动物以植物为食，有些则以其它动物为食，但所有动物最终都要依赖植物提供食粮」。

在 13 岁组别，增多了以下的一个概念：

- 绿色植物利用来自太阳的能量透过光合作用制造食物。

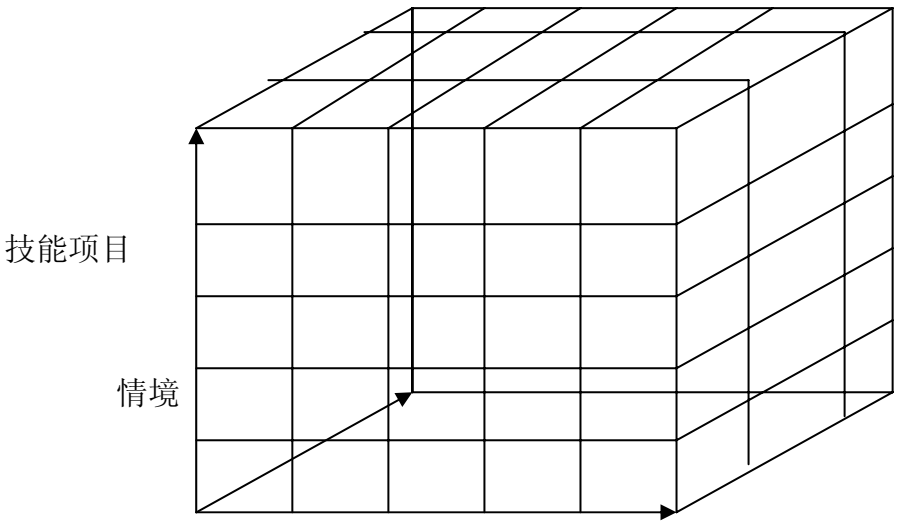
到 15 岁组别，再增添四个概念：· 在一个食物网中，如果任何一部份出现变动，其它部份都可能受影响；

- 食物网出现变化的原因可能是消费者和生产者之间的平衡状态发生了变化，或是因无机环境发生变化而造成；
- 生物生活于群落之中，而在其所属群落中，每一生物都占有一个最适于其生存的位置；
- 竞争和捕食有助维持一个群落中各个种群数量的平衡。

考题的情境

此外，小组亦希望监察学生在整个学校课程中的科学成就，而不单只是科学科上的表现(APU, 1979)。因此在拟题时融入了三类不同的情境：一是与科学课有关的，二是与其它科目有关的，而三是与日常生活有关的。例如，要评估考生是否懂得利用绝缘的方法去保温，会要求考生研究不同物质的导热性(与科学课有关的情境)，或指出未发明电冰箱之前，人们是利用甚么方法令东西保持清凉(与历史科有关的情境)，又或是找出保持家居温暖的方法(与日常生活有关的情境) (APU, 1979, p.5)。

基于以上所说，每一考题都包涵了三个向度：过程技能，内容范畴，及问题情境。这三个向度可以组合成一个三维框架(APU, 1979, p.4)，作为制订考题的基本指引；这框架亦有助于描述每一考题在整个评量中所发挥到的作用。



虽然小组有意透过是次评估，深入了解学生在横跨不同概念范畴及情境下的技能成就表现，但如果按照这个三维框架的每一个方格的要求拟题，题目的种类便非常庞大，处理上亦异常复杂。因此工作小组决定放弃以情境来将考题分类(DES, 1989)，但仍在考题中渗入不同情境，以保证评估的有效度。小组亦决定除了第四个有关应用相关科学知识的技能项目外，其它考题不应以学生对任何一项内容范畴的认识作为答题的必需条件。

调查对象及取样方法

根据 DES(1989)，是次评估的对象是 11，13 和 15 岁就读于英格兰，韦尔斯和北爱尔兰的学生。第一次大型取样调查始于 1980 年，以后每年进行一次，直至 1984 年完成最后的一次。调查的目的是监察每组的平均表现，及这表现是否会随时间而改变。在每个调查年度，每个年龄组别都有约来自 500 至 1,000 间学校的 12,000 至 16,000 名学童参与。在首年的调查，每所学校派出 11 岁学生 8 名和 13 和 15 岁学生各 9 名参予评估。抽样的方法是挑选每月首日出生的学生为评估对象。在其后的调查，每个年龄组别的参与学生的数目都有所不同。由于学生和考题都是以随机方式抽样，因此两方面的样本都应具有代表性。

试卷的设计

测考模式

由于评估所牵涉的内容项目众多，每次测试只能评估部份项目。每个子项目都设有一个试题库，部份试题适用于两个或三个年龄组别，以便于比较跨年龄的成就差异。每个子项目的试题库有约二百条考题，而每次评估测试只使用 45 至 60 题不等，然后再以随机抽样方式，抽取约 15 至 20 题印成试卷予学生作答。每次调查约使用 30 份不同的试卷。11 岁组别的评估时间约 45 分钟，而其它组别的时间为一小时。在笔试前或后，会抽取部份考生进行实作测试。

大部份考题是以笔试方式进行，部份项目例如使用量度仪器，及进行观察，则采用实作测试方式。在实作测试，每题所需的仪器会被放置于考场的一个区域内，考生轮流到不同场区作答(Schofield et al, 1982)。 在评估进行探究这主项时，除了根据考生的书面答案外，亦会参考观察员对考生的反应所作的记录。以书面作答的题目类型包括多项选择题，短答题及开放性的问 题。多项选择题的好处是可以缩短作答的时间及利用机器评分，也可以避免因评卷员的错误判断而出现误差，但缺点是可以让考生猜测答案。

实作测试有两种执行模式，第一种是循环方式，各考生在考场内轮流进行多项实作活动，包括运用仪器量度及观察，器材由大会向考生提供，虽然是实作测试，但考 生须以书面方式作答。第二种是个别测试，适用于评量第六主项，即进行整项探究。个别考生须单独接受评估员测试，评估员先向考生提出问题，引导考生进行实作 探究，评估员会在旁观察并记录考生所用的方法，事后评估员需要填写一份评估清单，以记录考生在探究中的表现，这种模式主要应用于进行探究这项目上，有关评 估实作测试的详细程序可参考 Welford et al (1985)，Murphy and Schofield (1984)，Murphy and Gott (1984) 及 Gott and Murphy (1987)。

每一考题的分数由零至三分，而多项选择题则为零至一分，由于选择题的分数较为划一，可以容许考评当局在每次评估时，采取较灵活的分层任意选题方式准备考卷。这样，每份考卷的试题组合便不相同，可以避免学校刻意操练学生应试，令评估更为公平。

学校问卷

此研究亦设有学校问卷，目的是收集各参与学校向学生提供的科学教育的资料，以便于找出学校所提供的科学教育与学生的成就表现之间的关联。问卷的内容包括科学教学的

师资和资源，进行科学教学的目的，对科学教育的重视程度，分配予科学教学部门的财政资源等。11 岁和其它年龄组别的问卷内容有很大差别，这是因为当时英国的小学所推行的科学教学存在颇大差异，但到了中学阶段，科学课程渐趋统一。以下是用于不同年龄组别的学校问卷的一些题目例子(Driver et al, 1982; Schofield et al, 1982)。

11 岁组别：

- 校内是由哪些人员负责科学教学？
- 科学课是否包含于校内学习活动之内？
- 进行科学教育是否学校的方针？
- 科学是时间表上编定的科目，还是占主题教学中的一个部份？
- 10/11 岁的学生用于学习科学的时间占全部课堂时间的百分之几？
- 科学活动有甚么目标？
- 学校对以下科学活动的重视程度。

  例如：

  - 小心地跟随着咭纸、书本或黑板上的指示去做
  - 在课堂上写下观察纪录
  - 制造一个良好的书面记录
  - 决定一个他们想调查的问题
  - 在真正测试之前估计测量读数
  - 小心进行直接的观察
  - 设计自己的实验

（问题经简化）

13 和 15 岁组别：

- 校内有多少名老师拥有正规的科学资历？
- 校内共有多少间实验室？
- 学校的科学部共有多少名实验室技术员？实验室技术员的工作时数有多少？
- 老师有没有使用校外资源，例如参观工业机构，到野外学习中心进行活动等？
- 在一个循环周内，共有多少堂科学课？
- 在学校的经费中，科学部所获得的分配有多少？

学生问卷

学生问卷的设置主要是为帮助分析学生的成就表现，以贯彻 APU 的第四项职能 — 识别学习环境与学生学业成就差异的关联。问卷旨在了解学生多方面的背景资料，例如：性别，种族，修读科目，职业取向，对有关科学议题的兴趣；问卷亦问及学生课余的兴趣及嗜好，这是基于越来越多研究发现，学生的课余活动与他们学习科学的兴趣有关。

效度和信度

为提高评估的效度和信度，小组邀请了科学教育的专家 包括大学教授，教师，考试局成员，督学等审议各范畴的主要项目，子项目和考题描述句，并试行将考题按照小组所设定的框架分类，从而评估框架的有效度，以及各主项目，子项目和考题描述句的清晰度，如有需要便提出修改，透过专家的共识，令评估框架更臻完善。

小组亦进行了先导试验，以分析考生在不同子项目上的表现的相关性。每一考题草拟后必需符合两个条件才会被纳入试题库之中，一是专家小组一致评定为具足够效 度；二是该考题的性质必需与其中一个考题描述句相符。此外，评分准则亦经标准化的过程。每位评卷员都是经过特别培训，以确保评分过程的可靠性。但如要保证 评量实践性考题的客观性及信度时，则较为困难。这方面惟有依赖先导试验，以减少由不同评估员作评量而出现的差异 (APU, 1979)。

对于进行整项探究的评量分析，小组针对其所设的探究框架下的每一个步骤拟出评量的仔细标准，然后统计达至这些标准的学生人数的百分比，以了解学生在不同步骤中的成就表现(Gott and Murphy, 1987)。

调查结果的分析

本调查结果分析可分为两个部份。第一部份是报告和分析学生的成就表现；第二部份分析学生的成就表现与各种背景因素的关联。

学生的成就表现

小组除了报告和分析学生在个别考题的表现外，还对学生在个别子项目，项目，及整体调查的表现作出概括性的分析；对个别考题的表现，报告列出了学生所达到的能力或分数的百分比；就每一子项目及项目，报告总结出各年龄学生的成就表现。以 13 岁组别于 1980-81 的调查为例，在利用图表及符号表达方式这项目之下的其中一个子项 - 利用图表及表格表达数据，很多学生对设定两轴的刻度感到困难。就这项目的整体表现而言，学生能够运用不同的数据表达方式，但如果学生同时要面对多种数据变化时，表现便较差；学生亦不大能将三维图像转译成正统的二维图形，例如线路图和切面图。这个年龄组别的整体调查果亦显示，学生在运用图表及表格，和应用生物概念这两个项目的成就存在很大差异，前者远较后者为优 (DES, 1984) 。

学生的成就表现与各种背景因素的关联

学校问卷和学生问卷所提供的资料可以用来分析影响学生成就的因素，以便于解释评量所发现的趋势。根据 DES (1989)的报告，这些因素包括学生年龄、性别、所修读的学科、班级人数、学生在校内接受教育的情况，以至学校对科学课程的安排，投放的资源等等。针对学生的因素而言，调查结果显示，在相同考题中，学生成就是随年龄而递升；学生的表现亦存在性别差异的情况。以应用物理学概念这子项为例，在三个年龄组别中，男生的表现都较女生为优；不过，在分析观察结果方面，却出现相反的情形。此外，调查亦显示男女差异的情况是与学校所属的地区有关，例如在观察和分析观察结果方面，韦

尔斯的 15 岁组别的男学生比女学生有较佳表现，来自英格兰和北爱尔兰的男生却刚好相反。综合各个项目的整体表现，北爱尔兰女学生的表现明显地比其它地区的女学生优胜。

针对学校因素而言，学生于子项中的表现与同类活动在学校的普及程度呈正相关。但有些结果却是出乎意料之外，例如，调查找不到证据显示，学生学习科学的时间越长，成就便越高；事实上，较平庸的学生学习科学的时间越长，应用概念的成就越低。此外，虽然位于近郊的学校所获分配的资源较少，但是这类学校的学生的表现却比城镇学校的学生为好。这些结果足以显示出学生的成就未必与学校因素有关。另一个出人意表的例子是在 13 岁学生中，班中的人数越多，学生的成就越高，这可能是因为学校早已将能力较差的学生编入小班之中，这个解释亦与学校的情况吻合。

以上例子说明了即使学生的成就与某种背景因素成正相关，也不能证实它们之间存在必然的因果关系。就正如调查发现修读物理科的学生比念生物科的表现较佳，并不一定表示这是修读不同科目的后果，而可能是因为成就较高的学生较倾向于拣选物理科而非生物科，当然小组亦不排除学生的表现可能是受到学科的不同教学方法所影响 (DES, 1989)。

出乎意料之外，学生在实作探究的表现比较他们在应用科学概念方面更为优胜，纵使课程偏重后者而忽略前者。根据小组的分析，学生在实作探究的表现可能是受五个因素所影响(Gott and Murphy, 1987)。第一是此类问题并不要求学生以文字作答，因此学生的表现不受语文障碍的影响。第二，此类实作探究提供了不少有关如何探究问题的提示，令学生不会感到问题难于理解，而缺乏此类提示正是文字题令学生感到困难的原因。第三，由于学生有机会不断尝试，他们可以根据结果而修订实验方法。第四，这些实作测试很少要求学生运用科学概念。第五，大部份实作测试是与日常生活有关，因而鼓励学生作出尝试。

调查的优点与不足

APU 可以说是开创了英国以技能为本的评量调查的先河，调查有系统地将科学探究对技能的要求有系统及具体地分为若干主项目和子项目，并以这些项目作为命题的基础，以保证考题的效度，成为日后此类调查的滥殇。调查亦将学生的成就表现与学生及学校的背景资料作比较，以辨析其中的关系所在，从而辨析一些可能影响学生成就的因素。

但是，本调查亦有很多不足之处，有些是与调查的目标和可行性有关，是关乎理论与实践的问题；有些是与问题或问卷的设计有关；亦有些是与其调查方法有关，关乎信度的问题。首先，APU 调查主要是希望监察在学儿童的成就，与及识别低学业成就的情况”，但调查的分析结果只能反映学生在本调查各项目和子项中的表现，而不能提供学生成就的标准或指标，因此很难对学童的整体成就作出判断。

第二，正如 DES(1989)指出，调查的评量框架主要反映参与调查计划的专家的共识，背后并没有很强的理论基础予以支持，所以难免会存在一些结构性的问题，例如一些子项之间互有重复的地方，包括“运用图表及表格” 与“分析数据，设计实验与进行实验等，

因此，学生在不同项目中的表现，并非完全独立。在拟订考题的过程中，子项目仍被不断修改，以避免重复，令人对这些评估结果的效度产生怀疑。

第三，小组亦承认过程技能的发挥，是很受探究的情境和学生对相关概念的掌握程度所影响，事实上，很多考题并不能将内容或概念完全抽离，例如在很多评估观察能力的考题中，只有那些已具备相关概念的考生，才能取得被评卷者视为正确的观察结果(Driver et al 1982)。 此外，很多子项目也要建基于数学及统计学概念例如比例，机率，平均值等 (DES, 1989) 。

正因如此，纵使学生在某些考题中有出色的表现，亦不一定表示他们能够将此等技能迁移至其它问题情境之中，好像在「进行探究」这项目中，学生在不同活动的表现存在显著差异，可见学生的成就表现是受其概念所影响。

第四，调查尝试找出与学生成就相关的因素，但从学校和学生问卷所搜集到的资料却未见深入，调查所得的相关因素研究结果与预期的有明显分歧，仍需要利用更深入的方法去了解不同相关因素之间的互动关系。

第五，由于每次调查所抽取的题目都稍有不同，因此隶属于相同子项目的考题的分数必须划一，否则该子项的总分便会出现不一致的情况。为了迁就分数的一致性，学生在个别考题所获得的分数，就未必能够完全反映他们在这些考题中的表现。

第六，此成就调查计划历时数载，考题也经历不少变化，例如某些题目原先只出现于一个组别，后来被应用于其它组别，由于考题已经历变化，所以很难比较在不同年份进行的调查的结果。此外，由于要符合不同组别的要求，同样的考题在不同组别的评分标准并非一致(DES, 1989)，因此如以调查结果来比较不同年龄的学生的成就或有欠公平。另外，一个更根本的问题是课程，教学方法甚至整个社会都不断改变，所以有需要按时更新考题，但这却与公平比较跨年级的进展背道而驰 (Gipps, 1988)。

第七，即使各年龄组别都使用相同的评估项目，但此方式有别于追踪研究，所以并不能用来了解学生的真正发展。

第八，正如 DES(1989)指出，任何大型测试都难免出现误差，所谓误差是指调查结果与真实情况之间的分别。调查的误差可以与取样有关或与取样无关。根据小组的分析，与取样有关的误差主要来自学生之间，学校之间，及问题之间的差异，前者包括能力上的差异，课程的影响，语文能力，对考题相关情境的熟悉程度等；而学校的差异则包括学校的课程，教学方法，学习氛围等方面的差异；考题的差异主要是来自考题的不同难度，研究结果的误差主要受样本的数量影响，一般来说，样本越大，出现的误差越小。

最后，虽然 APU 由始至终强调的是学生的过程技能，但其对学生其它方面的成就表现如概念的理解，及科学态度的忽视却成为了外界对它批评的原因，尤其是当时的课程甚少强调过程技能。

影响及启示

在英国 APU 调查开创了评量科学过程技能的先河，它的影响可以分为短期与长期两方面。短期来说，它加深了教师对科学成就评量的了解，及提供了很多在课堂上评量学生成就的意念和素材(Nuttall 1992)。虽然调查未能确实地辨析出影响学生成就的相关因素，但是，它至少反映出一些课程或学习的潜在问题，例如，学生应用科学概念的能力偏低，值得作进一步研究。长远来说，APU 将科学过程技能的评量具体化及系统化，影响了日后对科学成就的评量，它衍生出 GCSE (General Certificate of Secondary Education)的实作评量部份，及后期用于评量国家课程(National Curriculum)的标准评量任务。由于 APU 只提出了一个评量探究技能的框架，调查结果只能反映不同项目或子项之间的相对表现，所以小组并没有将各项目的表现转译成一个总分数，并以此标示每名学生或整个组别的成就表现(DES 1989)。因此，调查结果未能显示出学生是否已达到一定的标准，亦未能指出这绝对或相对标准应如何界定。不过，调查结果的公布开始令外界关注到应否为不同年龄的学生厘订成就标准。因为如果没有订出统一的标准，课程发展者便很难厘定清晰的目标，以配合不同年龄组别的发展。基于此，在九零年代初英国发展了国家科学课程，其中一个崭新目标就是发展学生的科学探究能力，课程石破天惊地提出了一套配合不同发展阶段的目标阶梯，作为学校发展课程和评量学生成就的准则， 这套准则的诞生与 APU 的调查不无关系，更可以理解为 APU 对科学探究所持有的理念的延伸(Donnelly and Jenkins, 2001)。该目标阶梯经多年的发展已趋于成熟，虽然还有未尽善的地方，亦未必能够全面及完全准确地反映学生在探究方面的能力发展，但它对于科学课程已起着两方面颇为深远的影响。其一是学校及教师需要以该目标阶梯为准则，设计针对不同年级或不同能力水平的课程，希望能够做到因材施教，从而逐步提升学生的水平。其二是让学校及教育当局以此作为评量各校各级学生的成就表现的划一准则。为配合国家课程的评鉴，英国评估当局发展了一套名为标准评量任务(Standard assessment tasks, SAT)的实作评量工具，以测试学生的探究技能，虽然后期基于教师工作量及时间上的限制，这套工具很大程度上被书面测试所取代，但毫无疑问，这套评量工具的影响已反映在课堂教学之上。由此可见，APU 所奠下的评估框架和理念还是有一定影响的。这方面的影响并不限于英国本土，还扩展及世界的其它地方。

参考书目

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.