課程與教學季刊　2010, 13(1), 頁 1～20

PISA 科學素養之試題認知成份分析

張銘秋* 謝秀月** 徐秋月***

認知取向的試題難度分析對教學與評量均有實質上的參考價值，本文整合素養架構與文獻，提出科學素養測驗之試題認知成份架構，描述科學素養測驗的試題難度與認知成份特徵，同時藉由不同國家的試題答對比率的差異討論，提供教學與評量的具體參考資訊。研究中所使用的資料庫是 PISA 2006 國際評量計畫中施測的 103 題科學試題。研究中以知識類別的數量、知識層次、科學能力與字數四個成份預測科學素養試題難度，此認知成份模式可預測約 52%的臺灣試題難度參數變異。在臺灣表現不佳的試題中，研究者提出不熟悉題型、不會使用題目給定的資料或證據形成推論、無法掌握變因以及無法從圖表中掌握答題的關鍵資訊等四項因素。此四項因素大多是因爲學生沒有機會接觸，不熟悉此種類型的試題。因此建議日後教師們可以增強這方面的教學。

關鍵字：科學素養、認知成份、國際評量

* 作者現職：國立臺南大學測驗統計研究所博士生

**作者現職：國立臺南大學材料科學系副教授

***作者現職：國立臺南大學測驗統計研究所助理教授

壹、緒論

一、研究動機

(一)認知成份分析在測驗發展的應用

傳統上，成就測驗的效度驗證都只集中在內容涵蓋性上，很少將焦點置於構念效度之上（Enright, Allen, & Kim, 1993）。在測驗架構或雙向細目表中包含認知或程序向度是越來越普遍了，雖然測驗架構或雙向細目引導測驗的發展，但它們並未直接接受實徵的驗證。測驗的認知分析是構念效度重要的論述依據之一，這個取向兼顧認知心理學和心理計量的模式（Embretson, 1993; Enright et al., 1993）。針對試題難度所提出的分析模式稱爲試題難度模式（Item Difficulty Models, IDM）（Gorin, 2005）。典型的 IDM 包含一系列解題運作歷程所涉及的認知處理或技能。IDM 的重要性在於確認試題處理的相關特徵，初步的試題特徵經常藉由內容領域相關文獻分析產生，實際運作的難度模式通常需要不斷反覆修正。

藉由 IDM 將這些認知成份區辨出來就可以直接進行教學。除此之外也可協助區辨現有架構與試題的缺點，提供不同形式試題與不同測驗間比較的基礎，對系統性的發展測驗頗有助益。再者，也可以協助測驗發展者在試題發展過程中，藉由操弄這些會影響試題難度的特徵，來估計試題的難度，因此不需要對每個試題都進行預試，既可節省成本，也可針對受試者的能力來配對試題難度。且透過這樣的分析也可以將非預期的試題難度揭露出來（是否有無關構念）（Dimitrov & Raykov, 2003; Enright et al., 1993; Rosca, 2004）。

(二)PISA 科學素養的內涵與趨勢

PISA 國際評量計畫(The Programme for International Student Assessment, 簡稱 PISA）是由經濟合作暨發展組織（Organisation for Economic Co-operation and Development，簡稱 OECD）所委託的評量計畫。PISA 測驗所評量的是十五歲學生分別在閱讀、數學、科學領域的素養成就，目的在於了解這些即將完成義務教育的學生，是否已經準備妥當成爲具有良好素養，且能積極貢獻社會的良好公民。

PISA 測驗的主要內涵並不在於鑑別學生是否有效地記憶已知的學科知識，而在於評量他們是否能夠把這些知識有效地應用，並且從不同角度分析與解決問題，應用於進入社會後所面臨的各種情境及挑戰，由此可見其重要性。臺灣在 2006 年的科學素養表現雖名列世界第四（林煥祥、劉盛忠、林素微、李暉，2008），但部分試題表現顯著低於世界一、二的芬蘭與香港，是急需加強之

處。

目前試題難度來源的研究多是針對數學（丁振豐，1997；劉子鍵、林世華、梁仁楷，1998；林世華、葉嘉惠，1999；Embretson & Gorin, 2001; Graf, Peterson, Steffen, Lawless, 2005; Katz, Lipps, & Trafton, 2002）或閱讀理解測驗（Breland, Lee, Najarian, & Muraki, 2004; Gorin, 2005; Gorin & Embretson, 2006）的試題，僅有部分是針對科學試題進行探究（涂柏原、梁恩琪、翁大德、楊毅立，2004；Enright et al., 1993; Rosca, 2004; Yepes-Baraya, 1996, 1997）。且上述研究所使用的測驗皆屬於成就測驗的範疇，尚未有針對素養（literacy）測驗的試題進行難度來源分析，因此本研究欲分析 PISA 2006 科學領域試題的難度來源，依據難度成份模式具體描述試題特徵，爲教學與評量提供更富認知意涵的參考資訊。

二、研究目的與研究問題

基於上述研究動機，本研究欲透過型塑試題心理特徵（characteristics）與觀察到的試題難度估計間的關係，發展 PISA 科學素養的認知成份模式，檢驗 PISA 科學領域試題的效度。研究目的分述如下：

(一)發展 PISA 科學素養的認知成份模式，了解各成份與試題難度之間的關係，與此成份模式對科學素養試題難度變異的解釋力。

(二)依據認知成份模式具體描述臺灣表現不佳的試題特徵，爲教學與評量提供更富認知意涵的參考資訊

貳、文獻探討

一、測驗認知分析的優勢

研究動機提及以認知取向爲依據來編製測驗是目前評量的趨勢，雖然現代測驗理論對測驗實務有許多實質上的影響，但心理計量和認知理論的連結性卻不大（Embretson, 1993）。認知心理學重視變異來源的操控，較具理論說服力，但其研究取向常忽視個別差異，且鮮少討論測量穩定性與精確性。心理計量學雖重視個別差異，在測量上講求穩定性及精確性，相對的卻缺乏效度的理論性驗證。因此結合兩取向的優點進行試題難度的分析，將有很大的學術產出性（Dimitorv & Raykov, 2003; Embretson & Gorin, 2001）。

Embretson 與 Gorin（2001）認爲以認知規則產生試題的研究趨勢，已從對能力的認知成份模式分析研究，擴展至使用成份來產生試題的應用模式階段。

認知系統設計是基於試題的特徵，預測試題心理計量特徵及反應時間的認知數學模式。試題的特徵與認知處理間是相關的，而認知處理可由難度的成份加以表示。一旦試題的特徵以數學模式建立，便可以在認知過程下操弄試題難度，藉以探討解體所需的認知能力。

除此之外，認知設計系統還具有以下的優勢：(1)因爲試題難度來源是由認知模式所解釋，因此可以預測新試題的參數；(2)試題的實徵特徵可由認知模式預測，因此不需對每個試題進行預試，可節省成本；(3)可禪明試題層次上的構念效度；(4)電腦產生試題不必依據現有試題，而是藉由特徵的重新組合來產生新試題；(5)可針對受試者的能力來配對試題難度；(6)可大量、快速的產生試題，有助於電腦化適性測驗的發展；(7)不需憂慮試題安全性問題，因爲系統中只有設計使用的認知因素是可知；(8)可針對認知及處理的特徵來發展教學策略（Dimitorv & Raykov, 2003; Embretson & Daniel, 2008; Embretson & Gorin, 2001）。

二、PISA 科學素養的內涵與評量架構

科學素養的評量，在學科部份包括物理、化學、生物及地球科學。試題類型可以分成三大類：(1)形成科學議題（Identify science issues）：要求學生從所提供的資訊之中，擬訂可以透過科學方法解決的研究問題；(2)科學化的解釋現象（Explain phenomena scientifically）：針對日常生活中常見的現象，如石雕受酸雨侵蝕，解釋其發生的原因；(3)科學舉證（Use scientific evidences）：利用科學證據來支持本身的主張或論點。

PISA 對於學生科學素養的能力，有以下幾個方面的界定：

個人擁有科學知識並能運用科學知識來界定問題、獲得新知識、科學化的解釋現象、並提出有證據支持的結論。

(一)能瞭解科學是人類知識與探究的一種形式。

(二)能意識到科學與科技是如何形塑物質、智力、與文化的環境。

(三)能成爲有反省力的社會公民，願意以科學的觀點投身於與科學相關的議題。

爲評量學生是否具備上述之科學素養，PISA 從四個彼此相關的向度著手：

(一)情境（context）：所謂情境是指包含了科學與科技的生活情境。

(二)知識（knowledge）：包括瞭解自然世界的科學知識與瞭解科學本身的

知識。

(三)能力（competencies）：形成科學議題（identify scientific issues）能力素養、科學化的解釋現象（explain phenomena scientifically）能力素養以及科學舉證（use scientific evidence）能力素養等三個項目。

(四)態度（attitudes）：意指對科學的興趣、對科學探究的支持、以及負責任地對諸如自然資源、自然環境等採取行動的動機。

架構圖如圖 1 所示。情境、知識、能力與態度四者彼此相關，並非相互獨立。這是因爲學生的知識、態度與情境脈絡無法脫離，而學生在情境脈絡中所表現的知識與態度，正是其能力的展現。

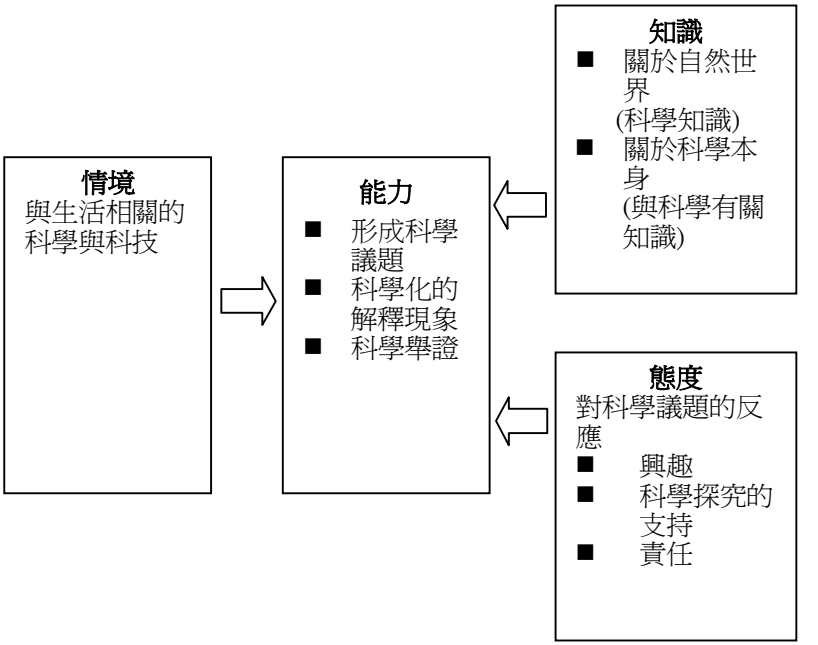

圖 1　PISA 2006 科學評量架構圖

三、科學領域的認知成份分析

由於本研究的焦點在於科學領域試題的認知成份分析，因此，此部分將焦點至於自然科學領域試題的難度分析研究。

Enright 等人（1993）以探討測驗的雙向細目表、專家看法與實際分析的方法，得到數個能預測科學成就測驗調查試題難度的來源。採用的是 44 個 1985-86 年 NAEP 生命科學分量尺試題的難度屬性。所使用的認知成份包含：

(一)文本與選項特徵：如段落/題幹及選項的字數或音節數、段落/題幹及選項的句數，若試題內有圖像材料，則這些圖像內的文字也會計算在內。

(二)試題的認知需求：以 Emmerich（1989）及 Scheumenan、Gerritz 與 Embretson（1991）所使用在 GRE 心理測驗試題上所發展出的認知需求，分爲 5 個層次：重述給定的訊息、區辨未給定訊息的正確部分、分析訊息、支持或削弱主張、程序或結果、綜合成份使其成爲一個新的組型。

(三)試題所需的知識層次：將每個試題歸到下列 6 個知識層級內：閱讀理解或問題狀態、日常生活、小學（k-3）、小學（4-6）、中等程度的（7-8）、進階的。但從編碼與檢驗資料發現「普通」不適用於 NEAP 試題，因此刪除此項知識層次。

(四)整體難度評估：由三位科學教學者評判每個試題對學生而言的整體難度。1 分表非常容易，5 分表非常困難。

Enright 等人（1993）發現科學教學者對整體試題難度的判斷對試題真實難度變異的解釋力爲 52%，是最有預測力的變項。而結合試題屬性訊息與教學者對試題難度的判斷，可以增加 7%-15%的解釋力。在整體難度預測之外，其他間斷的試題屬性對試題真實難度的解釋力也是研究的重點之一。結果顯示最佳的模式對難度變異的解釋力約爲 53%，所包含的屬性包括：知識層次、文本與選項組特質，而認知需求則不包含在模式內。其中，知識層次的獨特貢獻性約爲 38%，是最重要的屬性。

Yepes-Baraya（1996）則以 1993 年科學預試的兩個題本，研究學生在接受 NAEP 科學測驗時所使用的認知程序。對象是 16 個不同科學精熟水準的 8 年級學生，學生在接受測驗後進行訪談（放聲思考想像第一次作答，說出解題時的想法）。所採用的試題屬性，分爲五大類：1.內容知識、2.推理與解釋、3.形成與考驗假設、4.處理圖表訊息、5.試題形式與閱讀難度。雖未經統計考驗，但作者採用分析學生作答反應的方式，顯示該測驗確實評量到了其欲評量的構念，也就是科學知識結構、推理與形成與考驗假設。

隔年 Yepes-Baraya（1997）再探討 1996 年 NEAP 科學領域的所有試題（4,8,12 年級共超過 500 題）的難度來源。採用的試題屬性比 1996 年的研究多了實作作業的處理技巧，共有 39 個屬性。商請 9 位科學專家依據內容領域內原則與推理

的知識進行編碼。編碼與區塊分析顯示，帶有內容與說明的推理是最重要的屬性，是三種不同區塊成功解題的關鍵。

Rosca（2004）分析 1999 年 TIMSS 科學領域的 104 題選擇題的試題難度來源，所使用 17 個難度因素包括：題幹中的字數、文本中的字母數、文本中的段落數、文本中的句子數、每一個段落中的句子數、每一個句子中的字數、每一個字中的字母數、Flesch reading ease（可讀性指標）、Flesch-Kincaid grade level（可讀性指標）、認知層次（Bloon 的認知層次）、圖表呈現、解答的字數、誘答項的平均字數、選項數、正確答案字數與誘答項平均字數的比例、TIMSS 的科學內容領域、TIMSS 科學的表現層次。

Rosca（2004）採用迴歸與線性羅吉斯模式（Linear Logistic Test Model, LLTM）進行分析，結果發現對 1999 年 TIMSS 科學領域選擇題而言只有 Flesch reading ease、誘答項的平均字數、字數比例、認知層次與圖表呈現，這 5 個因素達顯著，但解釋力僅有 30%。

涂柏原、梁恩琪、翁大德與楊毅立（2004）針對 90 到 92 年度，六次的國中基本學力測驗自然科共 345 題進行試題分析，並提出下列五個難度因素：

(一)試題資訊量：試題中所提供成功解題之線索多寡，線索越多將越容易。

(二)選項異質性：選項內容不同，顯示同一題幹之下，學生欲成功解題所需具備的知識較多元，因爲每一個選項可能需要不同的知識或是概念，才可能順利作答；選項越異質，所需具備的基礎知識越多。

(三)圖文推測程度：指的是需從試題所提供的圖表等線索，推測判斷正確答案之程度，所需要推測的程度越高，試題難度將越難。

(四)轉化程度：試題若直接問一觀念，學生只需將正確知識表達即可，但若試題以轉化形式出現，則表示需要多一層思考運用，才能成功解題。

(五)知識量：試題本身所需具備的概念基礎多寡，將影響試題難度，所需概念基礎越多，試題難度將越難。

Scheumeman 等人（1991）則探討知識因素對成就測驗難度的影響。他們使用複雜因素去計算 GRE 心理測驗（知識特定的測驗）的試題難度。除了結構化的特徵之外，他們還使用了認知處理需求及所需的知識類別及層次。正確解答所需知識層次是由研究者來區辨，共分爲 5 個類別：閱讀理解（reading comprehension）、日常生活的（popular）、基礎的（basic）、中等程度的（intermediate）與進階的（advanced）。知識類別則包含：理論（theory）、準則

（criterion）、程序（procedure）、與關係（relationship）。

多元迴歸顯示對 GRE 心理測驗（高知識需求層次的測驗）難度的解釋力約爲 65%（變項包括：知識層次、知識類別、文本結構與 8 個可讀性指標或密度），其中最重要的因素是知識層次，獨特解釋力爲 21/%與試題所測量的知識類別。而對 NTE 溝通技能測驗（低知識需求層次的測驗）的試題而言則是命題數量與一個文本屬性因素，解釋力在 65%~68%之間。

叁、研究方法

本研究欲探討 PISA 2006 科學領域試題的認知成份模式，以下就資料來源與認知成份加以說明。

一、資料來源

本研究以 PISA 2006 素養作答反應的資料庫作爲資料來源。該資料庫中包含科學、數學與閱讀的原始作答反應，本研究僅採用其中科學領域的試題以及臺灣學生的表現進行分析。臺灣共有 8,815 位學生參與 PISA 2006 正式施測，PISA 雖以平衡不完全區塊（balanced incomplete block, BIB）設計，但 2006 年以科學爲主要施測科目，故 8,815 位學生皆有科學領域的作答反應。在試題部分 PISA 2006 資料庫中科學領域試題共 103 題，其中選擇題爲 37 題，複選題 32 題，開放式反應題 34 題，以 ConQuest 2.0 軟體，採用單參數模式估計試題難度參數。

二、認知成份及編碼說明

由於 PISA 科學試題並非只有選擇題的題型，因此與選項有關的認知成份不適用於本研究，再者，前述研究以多個變項來表徵閱讀負荷量，但多數都沒有顯著的解釋力，因此在本研究中僅單純的以字數來表徵閱讀負荷量。本研究所採用的認知成份分爲字數、認知需求及知識層次與知識類別數量，依序介紹如下：

(一)字數：

即文本、題幹與選項字數的總數，若題目標含圖、表，則圖表內的字數也列入計算。PISA 2006 科學領域試題的平均字數約爲 295 個字，因此以 150 字、300 字爲切點，將題目字數分成 3 個層次，在說明及解釋上與其它成份較爲一致。

(二)認知需求：

採用的是 PISA 的科學能力（competencies）：包括形成科學議題、科學舉證與科學化的解釋現象。

(三)知識層次：

採用 Enright 等人（1993）和 Scheumeman 等人（1991）的分類方式，但其中「日常生活」不適用於 PISA 的試題，因此不予採用。知識層次分爲閱讀理解或問題狀態、小學（K-3）、小學（4-6）、中等程度的（7-8）、進階的五個層次，各層次內容說明如下：

1.閱讀理解或問題狀態：所有所需的一般科學知識都由試題段落提供，所以會讓材料或問題更容易理解。

2.小學（K-3）：大部分學生第一次接觸到這種解題必須知識是在小學前期（幼稚園到小學 3 年級）。

3.小學(4-6):大部分學生第一次接觸到這種解題必須知識是在小學的 4-6 年級。

4.中等程度的（7-8）：大部分學生第一次接觸到這種解題必須知識是在中學的 7-8 年級。

5.進階的：試題要求要有更多進階的概念、更特定的知識細節或比前述階層更深入的理解。

(四)知識類別數量：

分爲四個類別分別爲：

1.**事實或理論的知識**：包括術語的知識、特定細節和元素的知識、分類和類別的知識、原理和通則的知識、理論模式/結構的知識。

2.**程序的知識**：包括特定學科的既能和演算知識、特定學科技術與方法知識以及運用規準的知識。

3.關係的知識：包括網絡、組型、序列、階層等系統性的關係，或因果、相關、影響等個別化的關係。

4.由實際經驗衍生而來的知識。

在編碼部分，知識類別並無層次之分，且 PISA 屬於素養的測驗，解題所

需的知識通常不侷限於一種。因此，知識類別數量此是檢視每一個試題解題時是否需要上述四種類別的知識，若需要則編爲 1，不需要則編爲 0，再將每一個試題在四個知識類別上的編碼加總，作爲知識類別數量的編碼結果。表 1 爲 4 種認知成份的代表說明。

表 1　PISA 科學領域試題認知成份代碼說明

代碼 / 成份	0	1	2	3	4
字數	150 字以下	151-300 字	301 字以上		
科學能力	科學化的解釋現象	科學舉證	形成科學議題		
知識層次	閱讀理解或問題狀態	小學(K-3)	小學(4-6)	中等程度(7-8)	進階
知識類別的數量	不需任何解題知識	需要一種解題知識	需要兩種解題知識		

以下以 PISA 科學領域範例試題及作爲編碼示例。PISA 範例試題及編碼如表 2 所示，茲說明如下：

範例試題一：試題 S485Q05 爲「酸雨」題組的問題之一，詢問學生在模擬酸雨對大理石的作用的實驗中蒸餾水的功用。試題原文：「學生們做這項實驗時，也放置一些大理石薄片在蒸餾水裡一整夜。請解釋學生們爲什麼在他們的實驗中包含了這個步驟。」

此題的總字數爲 153 字，因此字數編碼爲 1；在科學能力中屬於科學化的解釋現象，因此編碼爲 0；在知識層次中對照組的功用多在七年級或八年級教授，因此知識層次編碼爲 3。在知識類別數量中此題需要瞭解蒸餾水是爲了和酸與大理石的化學反應比較，所用到的知識爲關係的知識。但研究者檢視評分規範發現此題除了需要指出比較的作用之外，還需寫出酸（醋）對此反應而言是必要的才能達到滿分，若未進一步解釋酸（醋）對此反應而言是必要的，只能達到部分分數，因此將知識類別數量編碼爲 2。

範例試題二：試題 S493Q03 爲「運動」題組的試題之一。此題屬複選題，試題原文爲：

當肌肉被運動時發生了什麼事情？請就各項陳述，圈出「是」或「否」。

當肌肉被運動時這情況會發生嗎？	**是或否？**
肌肉獲得血液流量的增加。	是/否
脂肪在肌肉中形成。	是/否

此題總字數爲 98，故編碼爲 0；科學能力屬科學舉證，因此編碼爲 2；在知識層次中要知道運動時肌肉的實際變化要到七年級才會教授，因此編碼爲 3。在知識類別數量中此題只需知道運動時肌肉的變化，所用到的知識爲關係的知識，故編碼爲 1。

表 2　試題認知成份編碼說明

題號	難度	字數	科學舉證	知識層次	知識類別數量
S485Q05	1.137	1	0	3	1
S493Q03	-.841	0	2	3	0

肆、研究結果與討論

本章主要是根據研究目的呈現資料分析結果，並加以討論之。本章共分兩部分，第一部分主要是分析討論認知成份編碼的一致性，以瞭解編碼的適切性；第二部分探討認知成份對科學領域試題的解釋力。

一、PISA 2006 科學領域試題認知成份編碼一致性

由於知識層次與知識類別數量的編碼會受到編碼者主觀判斷的影響，導致編碼誤差的存在；因此，本研究採用百分比一致性來探討編碼的信度，以估計認知需求與知識層次兩個認知成份編碼的一致性。

由於大型測驗需建立題庫，因此 PISA 試題多數不公開。但每次正式施測後會有部分試題釋出，因此研究者以 2006 年正式施測釋出的 26 題範例試題（released item）與 27 題 2006 年以前就已經釋出的試題，共 53 個試題探討編碼一致性。在知識類別數量部分主要由臺南大學自然科學教育領域專家，與現任國中自然科教師擔任評分者。由於自然科學教育專家並未涉及國中、小自然領域的教學，對課程可能不夠熟悉，因此在知識層次的編碼上僅由研究者與該

現任國中自然科教師擔任。

研究者以三位編碼者兩兩配對在兩項認知成份編碼之差異進行評定者一致性百分比分析。由於試題數爲 53 題，因此共有 53x2=106 個原始反應，再由這些原始反應中找出評分完全一致的部分以計算評定者的百分比一致性。相較於 Enright、Allen 與 Kim（1993）的 50%~79%，本研究的一致性百分比（如表 3 所示）皆在 85%以上，顯示認知成份評定的共識建立不難。

表 3　認知成份編碼一致性百分比

評分者	認知成份　知識類別數量（完全一致）	知識層次（完全一致）
1 vs 2	94 （89%）	98 （92%）
1 vs 3	90 （85%）	--
2 vs 3	94 （89%）	--

二、試題認知成份對難度的預測力

表 4 爲 PISA 2006 科學領域試題各認知成份的描述統計值，由表中可看出 PISA 2006 科學試題的平均字數爲 294.91 個字，顯示其閱讀負荷量略重。研究者收集國民中學學生基本學力測驗 90 年、91 年與 92 年三個年度，6 次測驗共 345 題自然科試題與 TIMSS 科學科公布的中文試題 70 題計算試題平均字數。國中基測自然科 6 次測驗試題的平均字數約爲 106.78 個字，TIMSS 科學科平均字數約爲 70.72 個字，相對於 PISA 科學領域的 294.91 個字，可知 PISA 試題的閱讀負荷相較於一般成就測驗而言重了許多。

表 4　PISA 2006 科學領域試題認知成份之描述統計摘要（題數=103）

變項	最小值	最大值	平均數	標準差
原始字數	38	564	294.91	120.096
字數	0	3	1.38	.702
科學能力	0	2	0.96	.727
知識層次	0	4	2.43	.925
知識類別的數量	1	2	1.40	.492

表 5 爲試題難度與認知成份間的相關矩陣，其中試題難度與知識類別的數量相關最高，其次是知識層次，而字數與難度間的相關也達顯著。由表 4 可知

知識層次與知識類別數量相關亦達.01 的顯著水準，在多元迴歸由於預測變項不只一個，若預測變項間的相關程度過高，不但變項間之間的概念區隔模糊，難以解釋之外，亦會因爲預測變項間的共變過高，造成預測變項與效標變項共變分析上的扭曲，稱爲多元共線性（multicollinearnality）。因此研究者以條件指標（conditional index, CI）進行整體迴歸模式的共線性診斷，CI 值愈高表示共線性愈嚴重，Belsley、Kuh 與 Welsch（1980，引自邱皓政，2002）認爲 CI 值在 30 以下，表示共線性問題緩和；30 至 100 間，表示迴歸模式具有中度至高度共線性；100 以上則表示嚴重的共線性。而由表 5 可以看出本研究所使用四個預測變項的 CI 值皆小於 30，顯示共線性的問題並不嚴重，不至影響預測變項對效標變項解釋力的估計。

表 5　PISA 2006 科學領域試題難度與認知成份間之相關矩陣（題數=103）

變項	字數	科學能力	知識層次	知識類別的數量
難度	.202*	.482**	.593**	.597**
字數		-.005	.108	.191
科學能力			.223*	.201
知識層次				.463**

*p<.05, **p<.01

以表 5 的認知成份針對試題難度進行多元迴歸預測。表 6 爲多元迴歸分析結果，四個認知成份皆達顯著，而 CI 值也都在 30 以下，顯示共線性問題緩和。而四個認知成份對試題難度調整後的解釋力爲 52%，接近於 Enright、Allen 與 Kim（1993）研究所得的 53%，但兩研究所使用的認知成份有所不同。Enright 等人使用的是知識層次、文本屬性、選項誘答力以及解答與誘答選項字數比例四個認知成份。由於本研究使用選擇題、複選題與開放性試題，無法使用選項誘答力以及解答與誘答選項字數比例這兩個認知成份。

相對於 Rosca（2004）使用 Flesch reading ease、誘答項的平均字數、解答與誘答項字數比例、認知層次與呈現圖片與否，對 TIMSS 科學領域試題難度的解釋力只有 30%，相較之下本研究的解釋力高出許多。表 7 爲多元迴歸預測模式係數摘要表，試題難度的迴歸預測方程式爲：

$$\hat{Y} = .578X_1 + .331X_2 + .133X_3 + .001X_4 - 2.182$$

其中知識類別數量爲最主要的認知成份，獨特解釋力約爲 36%，其次爲知識層次。在 Enright 等人（1993）的研究中，知識層次的獨特貢獻性約爲 38%，是最重要的認知成份。因爲 Enright 等人所使用的是 NEAP 的科學領域，NEAP 屬於測量所獲得知識的成就測驗，因此知識層次是最重要的認知成份並非是意外的結果。而本研究所使用的 PISA 科學領域試題屬於素養的測驗，主要內涵並不在於鑑別學生是否有效地記憶已知的學科知識，而在於評量他們是否能夠把這些知識有效地應用，並且從不同角度分析與解決問題，應用於進入社會後所面臨的各種情境及挑戰。因此 PISA 解題所需的知識通常不侷限於一種。故在本研究中最重要的屬性是知識類別數量。但知識層次在本研究中的解釋力與知識類別數量幾乎沒有差異，顯示知識層次亦是重要的成份。

文獻顯示字數對試題難度並沒有顯著的預測力，但在本研究中字數亦具有顯著的預測力。研究者認爲這是因爲在 PISA 科學領域的測驗中試題的閱讀負荷量較一般成就測驗爲重，從表 4 可知 PISA 科學領域試題平均字數約爲 295 個字前述提及國中基測自然科試題平均字數約 107 個字，而 TIMSS 科學科的平均字數約爲 71 個字，顯示 PISA 科學領域的閱讀負荷量確實較大。換言之，試題的閱讀負荷量越大，將對學生構成越大的認知負荷。

表 6　難度參數多元迴歸預測方程式係數摘要表

預測變項	效標變項：PISA 2006 科學領域試題難度				
	非標準化係數	標準誤	標準化係數(Beta)	t	條件指數
(常數)	-2.182	.236		-9.248***	1.000
知識類別的數量	.578	.077	.344	4.297***	6.469
知識層次	.331	.148	.320	3.896***	3.787
科學能力	.133	.048	.217	2.778**	8.535
字數	.001	.001	.152	2.154*	9.609

*p<.05, **p<.01, **p<.001

三、臺灣表現較差試題特徵描述

探討造成臺灣學生表現較差的試題特徵，以提供教學與評量上的建議。研究者比較臺灣與科學表現位居世界一、二的芬蘭、香港與鄰近的韓國在個別試題上的表現差異。挑選出單題得分與三個國家中任一國家差異在.2 以上的試題，共 18 題。在這 18 題中臺灣的平均得分爲.46，芬蘭爲.73，香港爲.58，韓國爲.53，連整體排名在我國之後的韓國，在這 18 題上的表現都優於我國，且遠遠落後芬蘭與香港。

表 7 是表現較差試題與其他試題在認知成份上的差異。這 18 個試題中，7 題屬科學化的解釋現象，我國科學化解釋現象的素養能力雖居世界第三，但在此部分仍顯著落後於芬蘭與香港。再者，在 PISA 2006 中科學舉證共 29 題，在這 18 題中有 7 題屬科學舉證，佔全部科學舉證試題的 24%，顯示臺灣學生可能是因爲不熟練對證據的使用，而導致表現不佳。

由表中可看出臺灣表現較差試題的總負荷爲 7.11 大於其他試題的 5.87。其中差異最大的是知識層次，表現較差試題的知識層次爲 3.39，對照表 1 得知是中等程度，也就是 7-8 年級所學，而其他試題則約是 4-6 年級所學。7-8 年級在我國爲國中階段，開始分科教授與科學有關的學科，在廣度及深度上都比國小自然科複雜許多，在學習上的落差，可能是學生表現不佳的原因。

表現較差試題所需使用的知識類別數量高於其他試題，顯示臺灣學生可能無法整合 2 種或以上的知識來解題。此外有部分試題如範例試題一，要求學生進一步解釋作用的必要性或者必須提出充分的證據或具體的說明。但有部分的學生認爲理由淺顯易見，因此只有證據而沒有說明，是造成表現不佳的原因之一。而在字數上表現較差試題與其他試題則沒有差異，顯示臺灣學生在此部分試題上表現不佳非由字數差異所造成。

表 7　PISA 2006 科學領域臺灣表現較差與其他試題特徵描述

試題	題數	字數	科學能力	知識層次	知識類別數量	總負荷
表現較差試題	18	1.39	.83	3.39	1.50	7.11
其他試題	85	1.38	.75	2.36	1.38	5.87

另外，研究者分析試題內容之後，整理出 4 個臺灣學生表現不佳的原因，作爲教師們在教學或評量上的參考。

(一)不熟悉試題形式：

在 PISA 的科學領域試題中，如前述範例試題二的複選題約佔三分之一。而在上述 18 個臺灣表現不佳的試題中，有 11 題是屬於此類型的題目。在此種試題中的每個小題都要作答，且每個小題都答對才能獲得分數。但在臺灣的測驗中，無論是課堂或大型測驗幾乎沒有這樣的題型，導致學生不熟悉這樣的作答方式。許多學生都只回答第一個小題，而忽略其他小題，導致無法獲取分數。

(二)不會使用題目給定的資料或證據形成推論：

此類試題相當於 PISA 科學能力中的科學舉證。此類試題通常會給定證據與結果，要求學生根據證據與結果進行推論。在臺灣學生的訓練當中，強調的是理論的瞭解或現象的說明，並不強調證據與結果間的推論。學生不會使用試題所給定的訊息或資料，只會用自己的知識來解題，是在此類試題上表現不佳的主要原因。

(三)無法掌握變因：

此類試題相當於 PISA 科學能力中的形成科學議題。此類試題必須知道題目所指的變因爲何、題目所指的議題可不可以透過研究來解答。在臺灣學生的訓練中，在進行實驗時也是遵循教科書或教師事先布置好的議題進行探究。因此學生缺乏自行產生研究議題的經驗，自然無法掌握變因。

(四)無法從圖表中掌握答題的關鍵資訊：

在 PISA 科學領域的試題中，有部分試題是需要擷取圖或表的訊息來解答。但學生經常無法區分主要訊息與次要細節，誤把次要細節當成答題關鍵，甚至依本身的知識或經驗來作答。

由上述分析可知，臺灣學生在部分試題上的表現差異，並不是知識量上的不同，而是因爲臺灣的學生對試題類型的不熟悉或缺乏訓練所致。

伍、結論與建議

區辨出對試題難度有影響的因素之後，對教師及測驗發展者都有很大的幫助。在教學部分教師們就可以知道問題出在哪裡，如是測驗技巧、思考技巧、閱讀能力或內容知識，就可以直接進行教學。對測驗發展者而言，可以經由這些因素來控制試題的難度使能更有系統性、原則性的發展測驗，並且確保不會出現所不欲的試題難度來源，讓測驗解釋更具意義以及能有更好的構念效度（Enright et al., 1993; Rosca, 2004）。

本研究由認知負荷的角度出發，針對 PISA 2006 科學領域試題，提出字數、科學能力、知識層次與知識類類別數量等四項認知成份，可解釋約 52%的試題難度變異。其中知識類別數量與試題難度關聯最高，當四成份並呈時，知識類別數量的標準化迴歸係數也最高，可見解題所需知識類別的數量的認知負荷不容小視。

文獻顯示試題的可讀性對難度亦有顯著的預測力（Rosca, 2004;

Scheumenan et al., 1991）。但由於中文並沒有可讀性指標，而在本研究中僅使用字數代表試題的閱讀負荷量，可能過於粗糙，且沒有考量到字彙的層次，因此研究者認爲若將試題所使用的字彙層次也納入考量，也許會得到更高的解釋力。

此外，亦可以受試者的表現來描述屬性。因爲測驗表現是受試者與問題間的交互作用而形成的結果，所以必須瞭解受試者在這些情境之下會使用的知識與技能。且受試者答錯試題的原因可能有所不同。同樣的，難度相同的試題其難度來源並不完全相同。描述這些對試題難度有影響的不同因素是評估測驗構念效度的重要方式。若測驗的目的在於描述或診斷受試者的特質或表現，則更仔細、深入的理解這些問題或表現特徵就會變的更爲重要。

在臺灣表現不佳的試題上，主要差異可能來自於國小與國中的學習落差以及無法整合知識所造成。另外內容分析也發現不熟悉題型、不會使用題目給定的資料或證據形成推論、無法掌握變因以及無法從圖表中掌握答題的關鍵資訊等四項因素是表現不佳的主要因素。而此四項因素大多是因爲學生沒有機會接觸，不熟悉此種類型的試題。因此建議日後教師們可以增強這方面的教學。不僅是強調事實或現象的解釋，更要強化學生對資料、證據的使用。或者是只設定情境，由學生提出可進行研究的議題，加強科學探究的能力。

參考文獻

丁振豐（1997）。認知分析與心理計量分析對解平衡桿問題的認知發展層次與解題運作成份測量之比較。**國立臺南師範學院初等教育學報**，**10**，81-125。

林世華、葉嘉惠（1999）。數字系列完成測驗試題認知成份分析之研究。**教育心理學報**，**31**(1)，139-165。

林煥祥、劉盛忠、林素微、李暉（2008）。**臺灣參加 PISA 2006 成果報告**。行政院國家科學委員專題研究成果報告（編號：NSC 95-2522-S-026-002）。花蓮：國立花蓮教育大學；高雄：國立高雄師範大學。

邱皓政（2002）。**量化研究與統計分析：SPSS 中文版視窗資料分析範例解析**。臺北：五南。

涂柏原、梁恩琪、翁大德、楊毅立（2004 年 3 月）。**國中生基本學力測驗自然科試題分析研究**。論文發表於國立臺南師範學院主辦「科技化測驗與能力

指標評量國際研討會」。臺南：國立臺南師範學院。

劉子鍵、林世華、梁仁楷（1998）。二度空間視覺化測驗之試題產生算則的驗證與修正。**教育心理學報，30**(1)，177-193。

Breland, H., Lee, T., Najarian, M., & Muraki, E. (2004). *An analysis of TOEFL CBT writing prompt difficulty and comparability for different gender groups* (Educational Testing Service Report No. ETS-RR-04-05). Princeton, NJ: Educational Testing Service.

Dimitorv, D. M., & Raykov, T. (2003). Validation of cognitive structures: A structural equation modeling approach. *Multivariate Behavioral Research, 38*(1), 1-23.

Embretson, S. E. (1993). Psychometric model for learning and cognitive process. In N. Frederiken, R. J. Mislevy, & I. I. Bejar (Eds.), *Test theory for a new generation of tests* (pp. 125-150). Hillsdale, NJ: Erlbaum.

Embretson, S. E., & Daniel, R. (2008). Understanding and quantifying cognitive complexity level in mathematical problem solving items. *Psychology Science Quarterly, 50*(3), 328-344.

Embretson, S. E., & Gorin, J. (2001). Improving construct validity with cognitive psychology principles. *Journal of Educational Measurement, 38*(4), 343-368.

Emmerich, W. (1989, November). *Appraising the cognitive features of subject tests* (Educational Testing Service Report No. ETS-RR-89-53). Princeton, NJ: Educational Testing Service.

Enright, M. K., Allen, N. L., & Kim, M. (1993). *A complexity analysis of items from a survey of academic achievement in the life sciences* (Educational Testing Service Report No. ETS-RR-93-18). Princeton, NJ: Educational Testing Service.

Gorin, J. S. (2005). Manipulating processing difficulty of reading comprehension questions: The feasibility of verbal item generation. *Journal of Educational Measurement, 42*(4), 351-373.

Gorin, J. S., & Embretson, S. E. (2006). Item difficulty modeling of paragraph comprehension items. *Applied Psychological Measurement, 30*(5), 394-411.

Graf, E. A., Peterson, S., Steffen, M., & Lawless, R. (2005). *Psychometric and cognitive analysis as a basic for the design and revision of quantitative item models* (Educational Testing Service Report No. ETS-RR-05-25). Princeton, NJ: Educational Testing Service.

Katz, I. R., Lipps, A. W., & Trafton, J. G. (2002). *Factor affecting difficulty in the generating examples item type* (Educational Testing Service Report No. ETS-RR-02-07). Princeton, NJ: Educational Testing Service.

Rosca, C. V. (2004). *What makes a science item difficulty? A study of TIMSS-R items using regression and the linear logistic test model.* Unpublished doctoral dissertation, Boston College, Boston.

Scheumeman, J., Gerritz, K., & Embretson, S. E. (1991). *Effect of prose complexity achievement test item difficulty* (Educational Testing Service Report No. ETS-RR-91-43). Princeton, NJ: Educational Testing Service.

Yepes-Baraya, M. (1996, April). *A cognitive study on the National Assessment of Educational Progress (NAEP) science assessment*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, New York.

Yepes-Baraya, M. (1997, March). *Lessons learned from the coding of item attributes for the 1996 National Assessment of Educational Progress (NAEP) science assessment: Grade 4 results*. Paper presented at the Annual Meeting of the National Council on Measurement in Education, Chicago.

A Cognitive Component Analysis for PISA Science Literacy

Ming-Chiu Chang* Hsiu-Yueh Hsieh**
Chiou-Yueh Shyu***

The item difficulty cognitive component analysis is useful for the development of curriculum and assessment. This study integrates the framework and the literature perspectives to propose and interpret the item cognitive component model for an assessment of science literacy. The 103 items of The Programme for International Student Assessment (PISA)--Science Literacy were used for this analysis. Four cognitive components were proposed to predict the item difficulty parameters: the number of knowledge category, knowledge level, science competencies, and number of words. The results suggest that cognitive components can predict about 52% of item difficulty variance. The implications of these results for items that students did not perform well and for teaching strategies were also discussed.

Keywords: science literacy, cognitive component, international assessment

*Ming-Chiu Chang, Doctoral student, Graduate Institute of Measurement and Statistic, National University of Tainan

**Hsiu-Yueh Hsieh, Associate Professor, Department of Materials Science, National University of Tainan

***Chiou-Yueh Shyu, Assistant Professor, Graduate Institute of Measurement and Statistics, National University of Tainan