Bulletin of Testing and Assessment

2012年6月　第十期

考試學刊

研究報告

回顧及評價

大學入學考試中心（臺灣）

2011 年版大學學系探索量表的編製

劉兆明[1] 簡茂發[2] 洪冬桂[3] 林幸台[2] 王思峰[1] 陳清平[4]

劉澄桂[3] 區雅倫[3] 侯陳美[3] 蔡佳燕[3]

輔仁大學[1] 臺灣師範大學[2] 大學入學考試中心[3] 臺灣首府大學[4]

摘要

2011 年版大學學系探索量表係以美國勞工部（US Department of Labor）建構的職涯資訊系統（Occupational Information Network）中的知識領域為分析架構，以知識重要性區辨國內大學學程。本量表歷經二期計畫研究案，於第一期研究計畫中確認學習知識重要性為本量表分析架構，經分析全國 76 所大學學習知識與技能重要性調查結果，顯示 33 種知識重要性對國內大學系群之區辨甚高，對教育部 22 個學門之正確區辨力達 74.2%，而對本中心 18 個學群之正確區辨力也達 70.9%。第二期研究計畫則對大學校系所屬學類及學習知識重要性評定再次確認，並檢驗 30 種知識領域重要性評定對學群與學類的區辨力。在 123 個學類與 18 個學群的 30 種知識領域重要性評定的再測信度方面，學類的再測信度係數介於.8～1.0 之間，學群則介於.97～.99 之間。30 種知識領域重要性對 18 個學群之正確區辨力為 79.6%，123 個學類之正確區辨力為 80.5%（強迫進入法），皆比第一波資料之預測力為高，顯現校系再次確認其填答結果的效果。最後，依據此分析架構確認本量表測驗結果的呈現方式，透過對測驗結

果的解釋與應用，協助受試者找到與自己興趣適配的「學群」與「學類」，且進一步了解這些學群與學類內的學系以及這些學系未來發展的情形。

關鍵詞：大學學系探索量表、興趣、知識領域重要性、大學學類

劉兆明，輔仁大學心理學系教授
簡茂發，臺灣師範大學教育學系名譽教授
洪冬桂，大學入學考試中心副主任
林幸台，臺灣師範大學特殊教育學系教授
王思峰，輔仁大學心理學系教授
陳清平，臺灣首府大學幼兒教育學系副教授
劉澄桂，大學入學考試中心高級專員
區雅倫，大學入學考試中心資深專家
侯陳美，大學入學考試中心專門委員
蔡佳燕，大學入學考試中心專員

The Construction of the 2011 Edition of College Study Interest Inventory

Chao-Ming Liu[1], Maw-Fa Chien[2], Tung-Kuei Hung[3], Hsin-Tai Lin[2], Sy-Feng Wang[1], Ching-Pin Chen[4], Chen-Quie Liu[3], Ya-Lun Ou[3], Chen-Mei Hou[3], Chia-Yen Tsai[3]

Fu Jen Catholic University[1], National Taiwan Normal University[2], College Entrance Examination Center[3], Taiwan Shoufu University[4]

Abstract

The structure of the 2011 edition of College Study Interest Inventory is based on the knowledge classification of the Occupational Information Network (O*NET) constructed by U.S. Department of Labor, providing a high knowledge importance ratings of discrimination on college programs. The inventory was completed in two stages. The first stage advocates the importance of knowledge and discipline specific skills for each academic department and analyzes the survey data from 76 universities in Taiwan. Results revealed that the importance ratings of 33 knowledge categories yield high discrimination for college departments in two classification systems of Taiwan. The accuracy rate for discrimination of the 22 academic disciplines of Ministry of Education was 74.2%, and that of the 18 academic domains of the CEEC (College Entrance Examination Center) was 70.9%. The second stage reconfirms both academic classifications of academic departments

and their knowledge importance ratings. This stage also represents the discrimination of the ratings in academic groups and categories. The test-retest reliability coefficients ranged from .8 to 1.0 for 123 academic categories, and from .97 to .99 for 18 academic groups. The accuracy for discrimination of the importance ratings for the 18 groups and for the 123 categories was 79.6% and 80.5% respectively. This result demonstrates higher predicting power than the first stage, which reflects the reconfirmation effects provided by the academic departments in the first stage. The final version of the inventory was developed based on this analysis framework and results, with the hope of helping students find suitable college departments that adhere to their interests, and understand more about the study contents and future prospects of respective departments.

Keywords: College Study Interest Inventory, Interest, Importance of Knowledge Domains, Academic Categories

---

Chao-Ming Liu, Professor, Department of Psychology, Fu Jen Catholic University
Maw-Fa Chien, Professor Emeritus, Department of Education, National Taiwan Normal University
Tung-Kuei Hung, Vice President, College Entrance Examination Center
Hsin-Tai Lin, Professor, Department of Special Education, National Taiwan Normal University
Sy-Feng Wang, Professor, Department of Psychology, Fu Jen Catholic University
Ching-Pin Chen, Professor, Department of Early Childhood Education, Taiwan Shoufu University
Chen-Quie Liu, Senior Staff Member, College Entrance Examination Center
Ya-Lun Ou, Senior Specialist, College Entrance Examination Center
Chen-Mei Hou, Senior Executive Officer, College Entrance Examination Center
Chia-Yen Tsai, Staff Member, College Entrance Examination Center

壹、前言

大學入學考試中心為協助高中學生的生涯探索與規劃，曾與國內的學者專家合作編製「大學入學考試中心大學學系探索量表」，此量表於 1998 年發行時，分為自然組與社會組兩種題本，題本題項包括下列四種內容：一、高中生活--高中生平時喜歡做的事、二、高中課程--高中課程標準中的課程名稱或學習內容、三、大學學習內容--大學學系的學習活動及內容、四、生涯發展--畢業後所從事的工作內容或環境。量表以學類圖為分類架構，排除科際整合學系，列出 27 個大學系群，測驗結果將學生的興趣連結至數個大學系群以供參考（區雅倫、張郁雯、劉兆明，1998）。

當初量表是以「系群」與「學系」為區分單位，為讓每個系群有足夠代表性的題項，因此題項相當多，故需採文理分題本施測，再加上時代的變遷與學系的急速發展，除「高中生平時喜歡做的事」、「高中課程學習的內容」部分題項的內容可能與目前學生生活的經驗有落差外，大學各學系的「學習內容及活動」、「畢業後可能從事的工作」可能也與現況有所不同，而有修訂或重新編製題項的必要。此外，若仍以學系為單位進行修訂，未來亦將趕不上學系發展與變化的速度，因此量表的結構亦須重新建構。

基於上述理由，本中心自 2009 年啟動量表的修訂工作。歷經二期的計畫案，於 2011 年完成。2011 年版本擷取美國勞工部（US Department of Labor）建構的職涯資訊系統（Occupational Information Network，簡稱 O*NET）33 種知識領域中的 30 種（參見附錄 1）做為量表架構。本版量表共有 30 個分量表。題項包括高中生日常生活可能做的事、大學各學系的學習內容及活動，以及畢業後可能從事的工作等三類。鑑於學系跨領域發展的趨勢，且為讓學生有機會進行更寬廣的探索，此版本文理不再分組。

貳、量表簡介

本量表係以美國勞工部建構的職涯資訊系統（Occupational Information Network, O*NET）為參考架構。透過學生對 30 種知識領域的興趣以及大學各學系對知識領域重要性評定之比較，幫助學生瞭解自己與大學各學系適配的程度，選擇合乎自己志趣的發展方向。

本量表每種知識領域各有 5 題，總計 150 題。量表內容包括高中生日常生活可能做的事、大學各學系的學習內容及活動，以及畢業後可能從事的工作等三類。鑑於學系跨領域發展的趨勢，且為讓學生有機會進行更寬廣的探索，本版量表文理不分組。作答時間不限，全班作答完畢約需 25 分鐘。施測時學生逐題在「非常喜歡」、「喜歡」、「不喜歡」、「非常不喜歡」等四個選項中擇一作答。使用答案卡者，須將答案卡寄至本中心，經讀卡、電腦程式計算之後，列印出學生測驗的結果以及班級總表，隨同測驗結果說明與補充資料寄送至學校。

參、量表分析架構之確認

一、確認學習知識重要性為量表分析架構

劉兆明等人（2010）於探索量表修訂與應用第一期研究計畫中，完成量表修訂之基本架構，除確認 O*NET 知識重要性與技能重要性，能正確區辨 898 個美國的教育機構學程（Classification of Instructional Program, CIP）至 28 個大學學群與 8 個技職系群（知識重要性區辨力分別為 89.5%、85.3%，技能重要性區辨力分別為 85.8%、88.8%）外，為確認 O*NET 知識重要性與技能重要性於國內大學學程的適用性，於 2009 年 9 月進行全國 76 所大學學習知識與技能重要性調查工作，分析所回收 1314 個校系組（回收百分比 83%）學習內容於 33 種知識與 35 種技能重要與否的資料後，顯示 33 種知識重要性亦可適用於國內大學系群（教育部 22 個學門與本中心 18 個學群）之區辨（其正確區辨力分

別為 74.2%與 70.9%），而 35 種技能重要性則較不適用（其正確區辨力分別為 56.7%與 49.5%）。

二、大學校系學習知識重要性評定之再確認及其區辨力分析

本研究於第一期計畫中，除確認 O*NET 知識重要性與技能重要性，能正確區辨 898 個美國的教育機構學程至 28 個大學學群，與 8 個技職系群外，透過所蒐集 76 所大學校院學系於大學學習知識與技能重要性之資料，亦確認了 O*NET 30 種知識重要性適用於國內系群之區辨。由於個別校系於大學學習知識領域重要性調查表之填答結果易受填答者主觀因素影響，本研究將大學校系按其課程與其在 30 種知識領域重要性評定的結果加以歸類，除可降低個別校系於此調查表之評量誤差外，亦可於高中階段提供與學生興趣適配的大方向供學生探索與參考，對其生涯之規劃將較有助益。

以下分述於第二期計畫中，所作大學校系所屬學類及學習知識重要性評定之再確認及其區辨力分析的結果。

（一）大學校系所屬學類及學習知識重要性評定之再確認

經分析第一期研究計畫所蒐集 76 所大學校院學系於大學學習知識與技能重要性之資料後，將原本中心興趣量表之學群、學類分類系統加以調整成 123 個學類，以使部分校系可跨學類。經效度檢核，顯示調整後之分類系統合乎周延、互斥、簡約與合理等分類原則。

本研究參考美國 O*NET 之資料，界定 123 個學類之定義，再加上 123 個學類 30 種知識重要性之平均值，於 2010 年 9~11 月進行大學各校系所屬學類，及其前一年所填學習知識重要性調查結果再確認工作。大學學習知識重要性調查結果確認表共郵寄 76 所大學校院 1276 校系，結果回收 1540 校系組（1521 學系），新填校系 284 個，不計單招之校院學系，回收百分比為 95.4%，高於 100 學年度的調查回收百分比 83%。

於第二次調查中，原 1276 所校系/組中，有 56 所校系（4.39%）要求更改

學類，新增之 284 所校系/組中，有 16 所校系（5.63%）要求更改其所屬學類。經過濾各校系的學類歸屬要求後，可歸納出下列兩大類型：1、校系所要求歸屬的學類不存在與 2、相同教育部系類碼下之少數校系要求變更學類（亦即相同系類碼下之校系，不足半數要求更動）。若校系更改學類歸屬之要求屬於此兩大類型者，原則上傾向不同意變更。如：某校交通管理科學系原歸屬於「運輸物流管理學類」，該學系要求變更為「電信管理學類」，電信管理學類不存在，故並不同意校系所提出之學類變更。又如：某校應用美術學系原歸屬於「美術學類」，但該系所主張應歸入「藝術與設計學類」中，又應用美術學系隸屬於教育部 2111（應用）美術學系之系類下，在此次調查中共有 13 所校系同屬於此一系類下，但卻只有該系要求變更學類，在此情形下，為避免拆解系類後，造成日後學類歸屬複雜化與其他相關之衍生問題，故學類歸屬仍依據教育部之系類架構不予拆解。然而基於校系對自身學系了解最深，原則上，尊重各校系要求變更「學類」之意見，但為保有學類內的一致性，仍會對該系所提出之要求進行檢核，以決定是否更動。

（二）大學校系 30 種知識領域重要性評定之信度

30 種知識領域重要性評定可分個別校系層面與學類、學群層面，信度係指再測信度。校系 30 種知識領域重要性評定二次調查間隔將近一年，部分校系由於更名、改制、停招、系所合併或其它因素已不復存在，也未參與第二次調查，因此估計信度時不予納入。在為期一年的縱貫性調查中，共有 1243 所校系同時參與兩次調查，其中 544 所校系（約 43.76%）二次資料填答完全相同，423 所校系（34.3%）再測信度係數高於.9，整體調查信度係數高達.93。

在 123 個學類與 18 個學群的 30 種知識領域重要性評定的再測信度方面，於 1521 所校系中，123 所校系跨學類（占 8.9%），1398 所校系屬單一學類，計算學類 30 種知識領域重要性平均分數時，排除跨學類之校系，只計算屬單一學類校系之分數。分析結果顯示學類的再測信度係數介於.8～1.0 之間，學群

則介於.97～.99 之間，顯示學類或學群於 30 種知識領域重要性的評定，兩次調查結果具有高度一致性。

（三）30 種知識領域重要性對 18 個學群與 123 個學類之區辨力

在量表發展初期，除確認 O*NET 知識重要性能正確區辨 898 個美國的教育機構學程至 28 個大學學群與 8 個技職系群外，於第一次大學調查之後亦確認 O*NET 知識重要性可適用於國內系群之區辨，目前國內主要的兩大學術課程分類系統為教育部22個學門與本中心18個學群。經第二次的大學調查之後，再以 30 種知識領域重要性的評定對調整後的大學學類與學群的區辨力再次加以檢驗。

本研究使用區別分析的主要目的在判定歸類的正確率，區別分析（discriminant analysis）是多變量分析中用於判斷樣本歸屬哪個分類組群的統計方法，它在研究對象分類已知的情況下將相似的樣本歸為一類。區別分析根據樣本資料推導出一個或一組區別（判別）函數，同時指定一種區別規則，用於確定待區別樣本的所屬類別，使錯判率最小。本研究區別分析的自變項為 30 種知識領域重要性，依變項則為群組變項，使用了兩套既有之學系分類體系：1、教育部之「中華民國教育程度及學科標準分類」96 年第四次修訂版；2、大學入學考試中心之學群-學類分類系統（王思峰、劉兆明，2012）。區別分析乃以自變項的多條線性迴歸函數[1]，區分樣本在依變項上的差異（分類），進而預測樣本的特性或行為會偏向那一群組類別（Warner,2008）。

表 1 為 30 種知識領域重要性對教育部 22 個學門與 138 個學類之區辨分析結果。不論採用何種分析模式，30 種知識領域重要性對教育部 138 個學類之正確預測力約為 78.5%（強迫進入法）、72.6%（逐步迴歸法），皆較第一次資料之 77.5%、63.6%高；另外，30 種知識領域重要性對教育部 22 個學門之正確

[1] 迴歸函數的個數= min(k-1, p)，其中 k 為依變項的群組數，p 為自變項數。以表 1 為例，當依變項為教育部學科標準分類之 23 個學門、自變項為 30 個知識領域重要性，則該次區別分析會有 22 條迴歸式。故表 1 六次分析中，共有 126 條迴歸式，篇幅有限，不在此一一列舉這些迴歸式。

預測力約為 78.3%（強迫進入法）、77.5%（逐步迴歸法），納入其他學門之校系樣本則稍降低為 77.6%、76.2%，但亦皆比第一次資料之預測力高。由於各校系所屬教育部學門學類並未異動，顯現校系再次確認其填答結果的效果。

表 2 為 30 種知識領域重要性對本中心 18 個學群與 123 個學類之區辨分析結果，正確區辨力學群為 79.6%、學類為 80.5%（強迫進入法）。不論採取何種方法（強迫進入法或逐步迴歸法）、或選取不同的校系樣本範圍（不排除樣本、排除跨學類校系、排除跨學類以及跨學群校系），30 種知識領域重要性對學群與學類的區別力大抵接近。與第一次的調查結果比較，30 種知識領域重要性對學類之正確預測力由 78.3%提高至 80.5%（強迫進入法），若採逐步迴歸法，則由 65.5%更大幅度提高至 75.4%，除顯現校系再確認其填答結果的效果外，亦顯現調整分類系統的效果。本中心 18 個學群與 123 個學類請見附錄 2 與附錄 3。

最後，圖 1 與圖 2 比較 30 種知識領域重要性對教育部與本中心分類之區辨力（採強迫進入法），如圖所示，約在 80%~85%範圍，大抵可說是沒有明顯差異。

表1　30種知識領域重要性對教育部學門與學類之區辨分析結果

樣本選擇範圍	方法	含其他學門				不含其他學門			第一次資料
		學門數	*N*	學門正確區辨力	學門數/學類數	*N*	學門正確區辨力	學類正確區辨力	
不排除	強迫進入	23	1519	77.6	22/137	1501	78.3	78.5	77.5
排除跨學類校系	強迫進入	23	1396	80.3	22/129	1388	80.6	80.6	
排除跨學類&跨學群校系	強迫進入	20	1070	85.1	19/98	1065	85.2	86.5	
不排除	逐步迴歸	23	1519	76.2	22/137	1501	77.5	72.6	63.6
排除跨學類校系	逐步迴歸	23	1396	78.9	22/129	1388	79.1	74.6	
排除跨學類&跨學群校系	逐步迴歸	20	1070	82.7	19/98	1065	83.5	79.2	

註：區別分析時事前機率採依組別大小計算、遺漏值以 1 填補。

表2　30種知識領域重要性對18個學群與123個學類之區辨分析結果

樣本選擇範圍	方法	學群數/學類數	*N*	學群正確區辨力	學類正確區辨力	第一次資料
不排除	強迫進入	18/123	1519	79.6	80.5	78.3
排除跨學類校系	強迫進入	18/123	1396	80.8	81.5	
排除跨學類&跨學群校系	強迫進入	18/85	1070	84.5	84.6	
不排除	逐步迴歸	18/123	1519	79.7	75.4	65.5
排除跨學類校系	逐步迴歸	18/123	1396	80.2	77.7	
排除跨學類&跨學群校系	逐步迴歸	18/85	1070	83.6	81.0	

註：區別分析時事前機率採依組別大小計算、遺漏值以 1 填補。

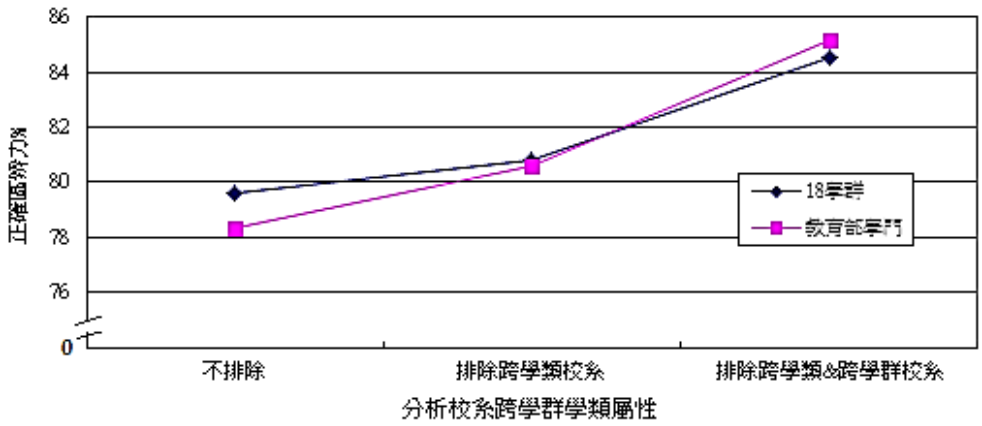


圖1　30種知識領域重要性對18個學群之區辨力

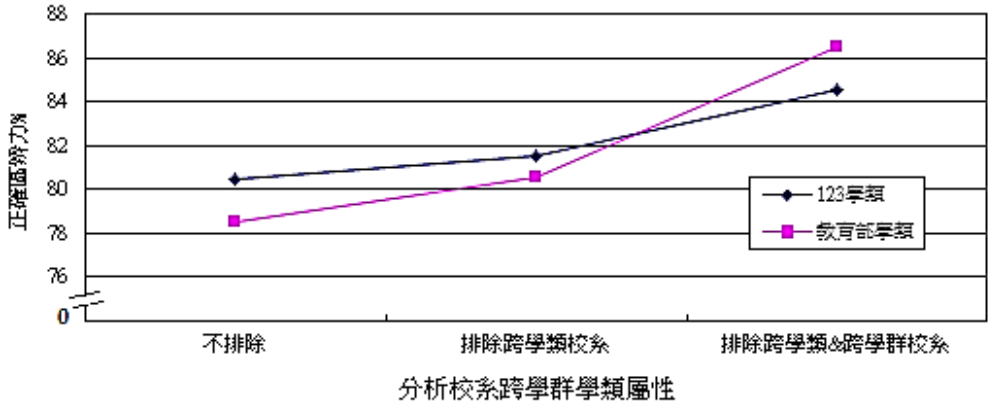


圖2　30種知識領域重要性對123個學類之區辨力

三、測驗結果連結學群、學類命中率分析

人境適配論一直是生涯發展理論的重要立論基礎，其前提是能用同一類別結構同時將人與環境予以區分成各類型，如此才能以此類別結構進行適配（Dawis & Lofquist,1984;Dunnette & Borman,1979; Holland,1973）。

本量表以 30 個知識領域為分析架構，以系所的知識領域重要性評定結果推論為該系所的學習特色心理特徵，以學生對知識領域感興趣程度的得分結果推論為個人的學習興趣心理特徵，依人境適配論的觀點，個人會傾向選擇與自己特質相似的環境，因此將系所的學習特色心理特徵與個人的學習興趣心理特徵，以 30 個知識領域作為串接架構，再以相關分析進行統計比較，進而推論出學生感興趣的大學校系。但由於大學校系數眾多，本量表測驗結果以大考中心的學系分類系統 18 個學群與 123 個學類作為分析單位，學生得到適配的學群與學類資料後，可透過大考中心網站上的「學系探索量表查詢系統」，搜尋各學群或學類所屬校系進一步的相關資料。

測驗結果要能有助於學生的生涯規劃或選擇校系，需與大學校系連結，而這種連結的方法與呈現方式需加以檢驗、確認。經分析第一期與第二期研究計畫中 76 所大學 1540 個校系組於學習知識領域重要性的評定分數與其所屬學類、學群之關係，以及 1460 位受測學生於 30 種知識領域得分之資料後，確定以相

關法連結學生測驗的結果至 8 個適配學類與 3 個適配學群，以及測驗結果的呈現方式。

首先將 1540 個大學校系組 30 種知識領域重要性評定分數與 123 個學類與 18 個學群於 30 種知識領域之分數求相關，校系組所屬第一學類為效標學類，校系組所屬第一學類之學群為效標學群。將各校系組與 123 個學類相關值或與 18 個學群相關值按大小排序，若效標學類相關值最高，則等第為 1，若相關值次高，則等第為 2，依此類推。效標學類、效標學群之相關值及其等第分布情形如表 3 及表 4 所示。以 r ≧.241（α＜.10）為切截點，分析結果顯示 96.7%校系組之效標學類相關值的等第出現在前 8 名，2.1%出現在前 9~16 名之間。同上分析，92.4%校系組之效標學群相關值的等第出現在前 3 名。

就命中機率而言，於 123 個學類中任取 8 個學類的機率為 8/123，其隨機命中率為.065（8/123），於 18 個學群中任取 3 個學群的機率為 3/18，其隨機命中率為.167（3/18），分析結果顯示，若以上述規則為各校系組取適配學類或學群，則取到效標學類或學群的命中率（.967、.924）遠高於隨機命中率。

表3　大學系組與效標學類之相關值及其等第分布情形（取前16個學類）

***N*=1540**

校系組數	百分比	取前16個學類累計命中率	效標學類相關值及其等第					
			效標學類相關值之等第	r ≧.463	.463＞r ≧.361	.361＞r ≧.306	.306＞r ≧.241	.241＞r
			顯著水準 α	0.005	0.01	0.05	0.1	
1413	91.8	91.8	1~4	✓				
1	.1	91.9	2		✓			
74	4.8	96.7	5~8	✓				
32	2.1	98.8	9~16	✓				
20	1.3		17 以上	✓				

表4　大學系組與效標學群之相關值及其等第分布情形（取前3個學群）

***N*=1540**

校系組數	百分比	取前 3 個學群累計命中率	效標學群相關值之等第	效標學群之相關值及其等第				
				r ≧.463	.463＞r ≧.361	.361＞r ≧.306	.306＞r ≧.241	.241＞r
1404	91.2	91.2	1~3	✓				
12	. 8	92.0	1~3		✓			
6	. 4	92.4	1~3			✓		
1	. 1	92.5	1~3				✓	
1	. 1	92.6						
			4 以上					
116	7.5			108	4	3	1	0

四、以相關法為學生取適配學群、學類之結果

同上，計算 1460 位學生於 30 個知識領域分量表的得分與 123 個學類、18 個學群 30 個知識領域重要性評定的相關，以上述規則為學生取適配學類與適配學群，結果如表 5 及表 6 所示。以 r ≧.241 為切截點，達標準的學類為適配學類，結果 93.6%學生（1367 人）可取到 9 個以上適配學類，5.7%（84 人）可取到 1~8 個適配學類，.6%（9 人）取不到適配學類。同此，以 r ≧.241 為切點，達標準的學群為適配學群，結果 73.9%（1079 人）可取到 4 個以上適配學群，19.8%（290 人）可取到 1~3 個適配學群，6.2%（91 人）取不到適配學群。進一步分析，顯示完全取不到適配學類與適配學群的學生占.6%（9 人）。

表5　學生取適配學類結果　　***N*=1460**

				學生得分與 123 個學類之相關值				
人數	百分比	累計百分比	達標準之學類數	r ≧.463	.463＞r ≧.361	.361＞r≧.306	.306＞r ≧.241	.241＞r
1190	81.5	81.5	17 以上	✓	✓	✓	✓	
67	4.6	86.1	17 以上		✓	✓	✓	
2	0.1	86.2	17 以上			✓	✓	
48	3.3	89.5	9~16	✓	✓	✓	✓	
51	3.5	93	9~16		✓	✓	✓	
8	.5	93.5	9~16			✓	✓	
1	.1	93.6	9~16				✓	
11	.8	94.4	1~8	✓	✓	✓	✓	
28	1.9	96.3	1~8		✓	✓	✓	
24	1.6	97.9	1~8			✓	✓	
21	1.4	99.3	1~8				✓	
9	0.6	99.9						無適配學類

表6　學生取適配學群結果　　***N*=1460**

				學生分數與 18 個學群之相關值				
人數	百分比	累計百分比	達標準之學群數	r ≧.463	.463＞r ≧.361	.361＞r≧.306	.306＞r ≧.241	.241＞r
951	65.1	65.1	4 以上	✓	✓	✓	✓	
102	7.0	72.1	4 以上		✓	✓	✓	
20	1.4	73.5	4 以上			✓	✓	
6	.4	73.9	4 以上				✓	
62	4.2	78.1	1~3	✓	✓	✓	✓	
103	7.1	85.2	1~3		✓	✓	✓	
63	4.3	89.5	1~3			✓	✓	
62	4.2	93.7	1~3				✓	
91	6.2	99.9						無適配學群

五、測驗結果呈現的方式

綜合上述研究結果，再徵詢學校輔導老師之意見後，確定測驗結果呈現的方式，內含測驗結果說明（30 種知識領域喜歡的程度、適配學群、適配學類、選系）、個別學生的測驗結果、大學學類網絡關係圖、30 種知識領域定義、學群及其主要學類以及學群描述等項目，以供實務上使用參考。

肆、測驗結果的解釋與應用

以大學學系探索量表測量學生對 30 種知識領域感興趣的程度，比對大學各學系對這 30 種知識領域重要性的評定結果，可以幫助學生瞭解自己與大學各學系適配的程度，進而朝向合乎自己志趣的方向發展。本量表可提供三種測

驗結果：30 種知識領域喜歡的程度、適配學群以及適配學類。量表的解釋包括剖面圖、適配學類與適配學群，量表的應用包括查詢系統及學類網絡關係圖。

一、量表的解釋

（一）剖面圖

學生對 30 種知識領域喜歡的程度同時以原始分數與百分等級呈現。每一種知識領域的原始分數最高 15 分，最低 0 分；原始分數愈高，表示學生對該知識領域喜歡的程度愈高。百分等級是將原始分數與全國高中學生比較，換算出學生在這方面高於多少人的百分比。如某生在某一知識領域的百分等級為 60，表示與全國高中生比較，某生對該知識領域喜歡的程度高於同領域 60%的人。

圖 3 剖面圖折線下的數字是某生在每一種知識領域的原始分數，條型圖上的數字是某生在每一種知識領域的百分等級。

（二）適配的學群

學生與 18 個學群是否適配，是將學生對 30 種知識領域喜歡程度的剖面圖與大學 18 個學群知識領域的剖面圖相比較，相關值愈高表示兩者愈相似。圖 4 條型圖上的數字就是學生與 18 個學群的相關值，0 以上為正相關，相關值愈高，表示與學生愈適配；0 以下則為負相關，負相關值的絕對值愈高，表示與該生愈不適配。

基本上我們選出相關值最高的前 3 個學群做為每一位學生最佳適配的學群。在選擇校系就讀或作生涯規劃時，學生可以優先考慮這些學群。圖 3 顯示某生最適配的學群是：遊憩與運動學群，管理學群，教育學群。

（三）適配的學類

18 個學群又可細分為 123 個學類，學生與 123 個學類適配情形的選定如同前述方法，即是將學生的剖面圖與 123 個學類的剖面圖相比較，相關值愈高表

示學生的剖面圖和該學類愈相似。再依適配程度列出與學生最適配的 8 個學類，較適配的 8 個學類，以及較弱適配的 8 個學類和最弱適配的 8 個學類，當學類未達標準時，適配學類數有可能少於 8 個。如前所述，某生的興趣與學類的適配情形如圖 5 所示：

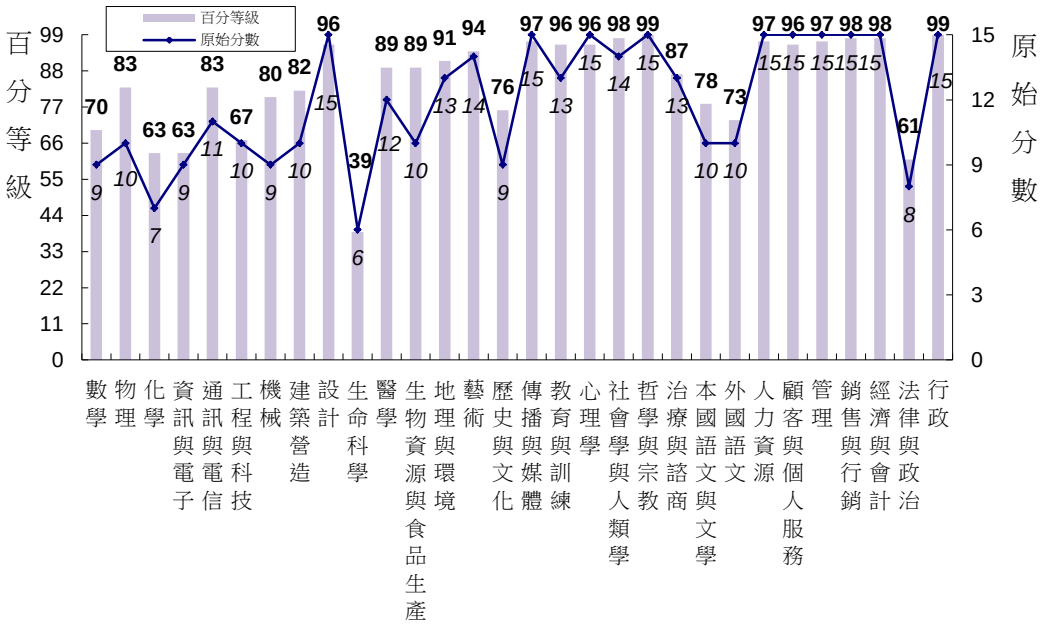


圖3　30種知識領域喜歡的程度之剖面圖(以某生為例)

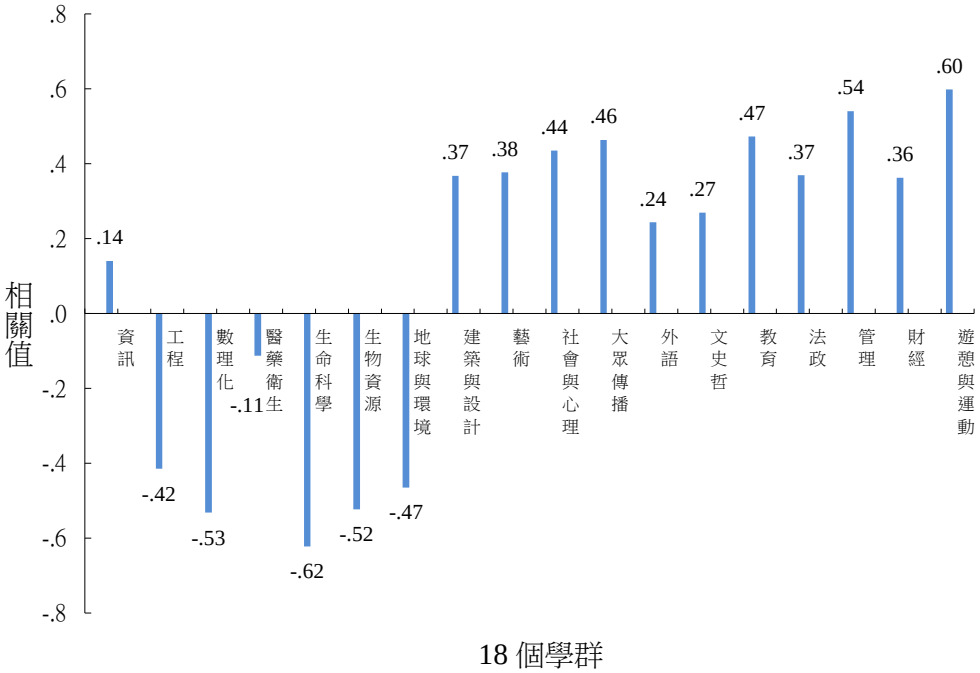


圖4　學生的興趣與18個學群的適配情形(以某生為例)

最適配學類							
(.71) 休閒管理	(.66) 成人教育	(.58) 運動管理	(.58) 社會工作	(.58) 舞蹈	(.57) 餐旅管理	(.56) 行銷經營	(.55) 觀光事業

較適配學類							
(.55) 教育	(.54) 數位學習	(.54) 時尚與服裝設計	(.54) 表演藝術	(.53) 廣告	(.52) 企業管理	(.50) 文化與產業	(.50) 圖書資訊

較弱適配學類							
(-.51) 農藝	(-.51) 光電	(-.53) 醫學影像暨放射科學	(-.53) 環境工程	(-.53) 獸醫	(-.54) 地質地球與環境科學	(.55) 植物保護與昆蟲	(-.55) 自然科學

最弱適配學類							
(-.56) 化學工程	(-.57) 生命科學	(-.59) 材料工程	(-.60) 生物科技	(-.62) 物理	(-.66) 生醫/醫學工程	(-.67) 化學	(-.68) 生物資訊

註:()內為你的 30 種知識領域分數與該學類 30 種知識領域重要性評定的相關值

圖5　學生的興趣與學類的適配情形(以某生為例)

二、量表的應用

（一）查詢系統

協助學生選擇校系就讀或作生涯規劃時，可以建議學生優先考慮最適配學

類，其次為較適配學類，同時也要慎重考慮避免選擇較弱和最弱適配的學類。至於學生的適配學群或學類內有哪些校系以及這些校系未來的發展為何，可請學生參考本中心網站「學系探索量表查詢系統」與「漫步在大學」。「學系探索量表查詢系統」內的資料除以上所述外，另包括 123 個學類的定義、該學類未來可從事的職業、該職業的 O*NET_Code、Holland 碼及從事該職業所需的學經歷等。

（二）大學學類網絡關係圖

除了學生的適配學群與適配學類內的校系外，還可以鼓勵學生進一步從知識領域的特性，探索與學生適配的學類，亦即以圖 6 的「大學學類網絡關係圖」進行生涯的探索。

圖中共有 123 個學類，虛線表示兩個學類有高相關。由於性質相近的學類在大學學類網絡關係圖中會座落於相近位置且會有虛線相連[2]，請學生將最適配學類在學類網絡關係圖中標註出來，這些學類都值得學生進一步探索與了解。

座落於圖的右半的學類主要屬自然科學方面；座落於圖的左半的學類主要屬人文與社會科學方面。亦有部分學類（如上方的心理學類及健康照護等學類，下方的資訊管理與統計等學類）位於自然科學、人文與社會科學之交界處，這些學類兼具文理性質。所有學類分屬三類組，圓形屬第一類組、方形屬第二類組、三角形屬第三類組。學類之形狀相同、深淺不同，表分屬不同學群。

有少數學生沒有得到適配學群或學類的建議，可能由於這些學生對多數的知識領域的喜歡程度都很相近，形成比較平直的折線圖，此時就會與任何一種學群或學類的折線圖都不相似，有這樣的情形時建議與學生一起探討可能的原因。

---

[2] 虛線表示兩個學類有一定的知識關連性，故標出該連帶(tie)，具體操作上，乃將兩個學類知識剖面之相關係數轉換為二分變項：小於 0.55 或負相關者為無連帶(0)、大於等於 0.55 者為有連帶(1)，並以虛線粗細表示其相關程度，愈粗者表示相關愈高。該圖乃以社會網絡分析軟體 Ucinet 所繪製(Borgatti, Everett & Freeman, 2002)，其空間位置乃由學類彼此關係連帶所界定，彼此關係連帶愈多者，即愈會安排在相鄰的空間位置，此與 MDS 的繪圖邏輯不同。

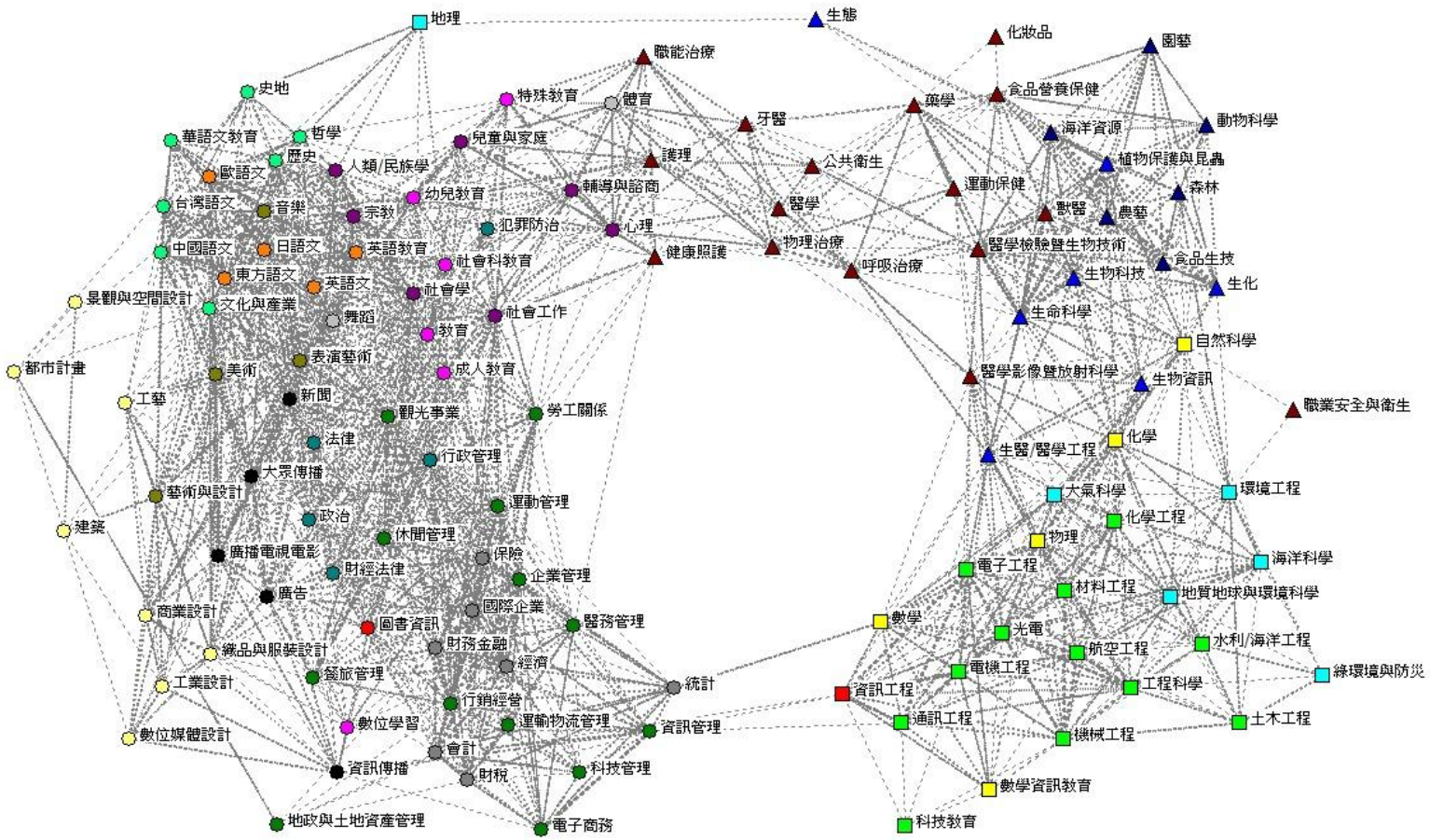


圖6　大學學類網絡關係圖

伍、結語

大學學系探索量表是為協助高中學生瞭解自己對大學各學系學習興趣而編製，它可幫助學生找到與自己興趣適配的「學群」與「學類」。本量表除可供高二下學生進行生涯探索與規劃或高三學生選擇校系參考外，本量表亦可作為一般及科技大學學生轉系與選擇研究所的輔助工具。然而，由於量表內容涉及大學各領域的學習知識，對於高一與高二上的學生較不適用。至於高職與科技大學學生以及成年人若考慮進一般大學就讀，也可用本量表作為生涯探索與規劃的輔助工具。

參考文獻

王思峰、劉兆明（2012）。學涯與職涯分類系統之串接：建立大學學系學類之關係描述子。**輔導與諮商學報**，**34**（1），1-29。

區雅倫、張郁雯、劉兆明（1998）。**大學入學考試中心大學學系探索量表使用手冊**。臺北：大學入學考試中心。

劉兆明、簡茂發、洪冬桂、林幸台、陳清平、王思峰、…區雅倫（2010）。**大學學系探索量表修訂與應用工作計畫結案報告**。臺北：大學入學考試中心。

劉兆明、簡茂發、洪冬桂、林幸台、陳清平、王思峰、…區雅倫（2011）。**大學學系探索量表修訂與應用工作計畫（二）結案報告**。臺北：大學入學考試中心。

Betz, N. E. (1987). Use of discriminant analysis in counseling psychology research. *Journal of Counseling Psychology*, *34*, 393-403.

Borgatti, S. P., Everett, M. G., & Freeman, L. C. (2002). *Ucinet for Windows: Software for social network analysis*. Harvard: Analytic Technologies, available in: http://www.analytictech.com.

Costanza, D. P., Fleishman, E. A., & Marshall-Mies, J. (1999). Knowledges. In N. G. Peterson, M. D. Mumford, W. C. Borman, P. R. Jeanneret, & E. A. Fleishman (Eds.), *An occupational information system for the 21st century: The development of O*NET* (pp. 71-90). Washington, DC: American Psychological Association.

Dawis, R. V., & Lofquist, L. H. (1984). *A psychological theory of work adjustment.* Minneapolis, MN: University of Minnesota.

Dunnette, M. D., & Borman, W. C. (1979). Personnel selection and classification systems. *Annual Review of Psychology*, *30*, 477-525.

Holland, J. L. (1973). *Making vocational choice: A theory of careers*. Englewood Cliffs, NJ: Prentice-Hall.

Jöreskog, K. G., & Sörbom, D. (1993). LISREL 8: *Structural equation modeling with the SIMPLIS command language*. Chicago, IL: Scientific Software International.

Peterson, N. G., Mumford, M. D., Borman, W. C., Jeanneret, P. R., Fleishman, E. A., Levin, K.Y., & Dye, D. M. (2001). Understanding work using the Occupational Information

Network (O*NET): Implications for practice and research. *Personnel Psychology, 54*, 451-492.

Prediger, D. J. (1982). The marriage between tests and career counseling: An intimate report. *The Vocational Guidance Quarterly*, *28*, 297-305.

Warner, R. M. (2008). *Applied statistics: From bivariate through multivariate techniqeus.* Thousand Oaks, CA: Sage.

Zytowski, D. G. (1994). Test and counseling: We are still married and living in discriminant analysis. *Measurement & Evaluation in Counseling & Development*, *26*, 219-223.

附錄一 30 種知識領域定義

1.數學	學習關於數學、統計學方面的知識及其應用。
2.物理	學習關於物理原理與定律、物理現象的解釋等方面的知識。
3.化學	學習關於化學元素、化學結構、及化學作用等方面的知識。
4.資訊與電子	學習關於電路板、晶片、電子設備、電腦軟、硬體方面的知識。
5.通訊與電信	學習關於電訊的傳輸、發射、交換等電信及通訊系統方面的知識。
6.工程與科技	學習應用工程學的專業與技術，設計與生產各種產品與服務的知識。
7.機械	學習有關機器、工具機等機械的設計、使用及維修的知識。
8.建築營造	學習關於建築物、橋樑、道路等之設計、建造及工法方面的知識。
9.設計	學習有關構思、藍圖、製圖和模型製作等的設計方法和技巧的知識。
10.生命科學	學習有關動、植物、微生物等有機體的細胞、基因、分子及生化作用的知識。
11.醫學	學習有關症狀診斷、疾病治療、藥物作用等方面的知識。
12.生物資源與食品生產	學習有關動植物養殖及食品生產、食品保存等技術與設備的知識。
13.地理與環境	學習有關地理學描述地貌、海洋、氣團、物產、及生態環境方面的知識。
14.藝術	學習有關音樂、舞蹈、視覺藝術、戲劇詩歌等的創作與表演的知識。
15.歷史與文化	學習有關歷史事件及其起因，對人類文明與文化影響的知識。
16.傳播與媒體	學習有關媒體製作、傳播技術和方法等方面的的知識。

17.教育與訓練	學習有關課程設計、教學方法及評量、教育訓練的原理和方法的知識。
18.心理學	學習有關人類的行為表現，能力及性格，動機與學習，異常行為等方面的知識。
19.社會學與人類學	學習有關社會群體行為及人際互動、社會發展趨勢、人類遷徙、種族、文化的知識。
20.哲學與宗教	學習有關哲學觀點、宗教教義、宇宙本質等及其對人類文化影響的知識。
21.治療與諮商	學習關於心理功能失常的診斷及治療的原理和方法，及生涯諮商等的知識。
22.本國語文與文學	學習有關本國語文結構與內容的知識，包括文字的意義、聲韻與寫作的知識。
23.外國語文	學習外國語文的結構與表達，包括文字意義、拼音、發音及文法等。
24.人力資源	學習有關人員招募、甄選及訓練的方法、程序，及福利津貼、勞資關係、協商及員工資訊等知識。
25.顧客與個人服務	學習有關認識市場、顧客需求及服務滿意度調查方面的知識。
26.管理	學習有關經營與管理的策略、領導技巧、生產效率、人力與資源協調的知識。
27.銷售與行銷	學習有關行銷策略、產品展示、銷售技巧及銷售控制方面的知識。
28.經濟與會計	學習有關經濟和會計事務的工作，金融市場、銀行業務及財務分析報告等知識。
29.法律與政治	學習關於法律、法庭程序、判例、政府法規、行政命令、公務規章等知識。
30.行政	學習關於機構的行政管理、人事行政、公共事務處理及危機管理等方面的知識。

附錄二 學群表

	學群名稱
1	資訊學群
2	工程學群
3	數理化學群
4	醫藥衛生學群
5	生命科學學群
6	生物資源學群
7	地球與環境學群
8	建築與設計學群
9	藝術學群
10	社會與心理學群
11	大眾傳播學群
12	外語學群
13	文史哲學群
14	教育學群
15	法政學群
16	管理學群
17	財經學群
18	遊憩與運動學群

附錄三 學類表

	學類名稱		學類名稱		學類名稱
1	資訊工程學類	42	森林學類	83	英語教育學類
2	電機工程學類	43	海洋資源學類	84	中國語文學類
3	光電學類	44	獸醫學類	85	歷史學類
4	電子工程學類	45	植物保護與昆蟲學類	86	哲學學類
5	通訊工程學類	46	生態學類	87	台灣語文學類
6	機械工程學類	47	食品生技學類	88	史地學類
7	航空工程學類	48	地質地球與環境科學學類	89	華語文教育學類
8	土木工程學類	49	地理學類	90	圖書資訊學類
9	水利/海洋工程學類	50	海洋科學學類	91	文化與產業學類
10	化學工程學類	51	大氣科學學類	92	教育學類
11	材料工程學類	52	綠環境與防災學類	93	特殊教育學類
12	生醫/醫學工程學類	53	環境工程學類	94	幼兒教育學類
13	工程科學學類	54	都市計畫學類	95	成人教育學類
14	科技教育學類	55	景觀與空間設計學類	96	社會科教育學類
15	數學學類	56	工業設計學類	97	數位學習學類
16	化學學類	57	工藝學類	98	法律學類
17	物理學類	58	商業設計學類	99	財經法律學類
18	自然科學學類	59	時尚與服裝設計學類	100	政治學類
19	數學資訊教育學類	60	藝術與設計學類	101	行政管理學類
20	醫學學類	61	數位媒體設計學類	102	企業管理學類
21	公共衛生學類	62	建築學類	103	行銷經營學類
22	牙醫學類	63	美術學類	104	運輸物流管理學類
23	物理治療學類	64	音樂學類	105	資訊管理學類
24	職能/語言治療學類	65	表演藝術學類	106	電子商務學類
25	護理學類	66	心理學類	107	科技管理學類
26	醫學檢驗暨生物技術學類	67	社會學學類	108	醫務管理學類
27	醫學影像暨放射科學學類	68	社會工作學類	109	勞工關係學類
28	藥學學類	69	人類/民族學學類	110	地政與土地資產管理學類
29	食品營養保健學類	70	兒童與家庭學類	111	觀光事業學類
30	呼吸治療學類	71	宗教學類	112	運動管理學類
31	健康照護學類	72	輔導與諮商學類	113	餐旅管理學類
32	化妝品學類	73	犯罪防治學類	114	休閒管理學類

33	職業安全與衛生學類	74	大眾傳播學類	115	會計學類
34	運動保健學類	75	廣播電視電影學類	116	財務金融學類
35	生命科學學類	76	新聞學類	117	國際企業學類
36	生物科技學類	77	廣告學類	118	財稅學類
37	生物資訊學類	78	資訊傳播學類	119	保險學類
38	生化學類	79	英語文學類	120	經濟學類
39	農藝學類	80	歐語文學類	121	統計學類
40	動物科學學類	81	日語文學類	122	體育學類
41	園藝學類	82	東方語文學類	123	舞蹈學類

高中職學生升學考試中的白話文閱讀選擇題組研究

——91～100 年試題的素材與閱讀層次分析

游適宏

國立臺灣科技大學

摘要

本文從「高中、高職升大學（含科技校院）」的大型入學測驗——學科能力測驗、指定科目考試、四技統一入學測驗中，選取 91 至 100 年間國文科的「白話文閱讀選擇題組」為研究對象，就其 46 個「閱讀素材」的種類與 108 個「試題」所涉及的閱讀層次進行分析，除了藉以了解國內閱讀能力檢測的部分現況，所得到的研究結果亦可做為閱讀教學與試題編製的參考。

關鍵詞：閱讀評量、國文試題、學科能力測驗、指定科目考試、四技統一入學測驗

游適宏，國立臺灣科技大學人文社會學科副教授

The Reading Tasks on Chinese Modern Essays of College Entrance Examination（2002-2011）

Shih-Hung You

National Taiwan University of Science and Technology

Abstract

This project consists two parts. The first part analyzed the structures of reading tasks on Chinese modern essays on the General Scholastic Ability Test, the Department Required Test, and the four-year TVE (The Technological and Vocational Education) college entrance examinations, including 46 texts and 108 items. The second part analyzed examinees' performances on different reading texts and aspects of reading.

Keywords: Reading Assessment, Chinese Test Items, The General Scholastic Ability Test, The Department Required Test, The Four-Year TVE (The Technological and Vocational Education) College Entrance Examination

---

Shih-Hung You, Associate Professor, Department of Humanities and Social Science, National Taiwan University of Science and Technology

壹、緒說

一、研究緣起

基於「閱讀力即國力」、「閱讀是教育的靈魂」，教育部自民國 93 年迄今，已連續推動「焦點三百：國小兒童閱讀計畫」、「偏遠地區國民中小學閱讀推廣計畫」、「悅讀 101：教育部國民中小學閱讀提升計畫」、「閱讀植根與空間改善：98—101 年圖書館創新服務發展計畫」，期能達到全民閱讀、終身學習的目標（吳清基，2010）。然而，臺灣在首度參加 PISA2006 閱讀、PIRLS2006 這兩項國際評量[1]後，表現卻不理想——八年級學生的閱讀能力在參與 PISA2006 的 57 個國家中排名第 16，落後韓國、香港，四年級學生的閱讀能力在參與 PIRLS2006 的 45 個國家中排名第 22（張莉慧，2009）。俟 PISA2009 閱讀評比結果出爐，臺灣更在兩岸三地敬陪末座——八年級學生的閱讀能力在 68 個參與國家中排名 23，遠落後第 1 名的上海及第 4 名的香港（天下雜誌教育基金會，2010）。這使臺灣的學校教育、考題型態能否培養閱讀能力，更加受到關切[2]。

中學教育的各種課程，於課程綱要中明揭以「提高閱讀能力」為目標者，大概只有國文，因此，國文科考試通常少不了閱讀能力的檢測。而全國性、且與升學相關的國文考試，由於參與人數眾多、試題編製嚴謹，不但可以從題目看到專家們設計的閱讀能力檢測方式，也可以從作答結果看到考生們的閱讀能

[1] PISA 是經濟合作發展組織（Organisation for Economic Cooperation and Development，OECD）所主持的「國際學生評量計畫」（The Programme for International Student Assessment），評量八年級（15 歲）學生在數學、自然科學及閱讀方面的能力。PIRLS 則是國際教育成就評價學會（International Association for the Evaluation of Educational Achievement，IEA）所主持的「促進國際閱讀素養研究」（Progress in International Reading Literacy Study），以四年級（9—10 歲）學生為施測對象。

[2] 例如 2010 年 12 月出刊的《親子天下》19 期，便特別企畫「為何少年不閱讀——走錯方向的國中教育」，有陳雅慧〈誰把臺灣少年考笨了〉、賓靜蓀〈PISA 啟示錄：走錯方向的語文教育〉等數篇報導，明確指出：「翻開 PISA 的測驗試題，任何一位關心教育的讀者，都可以立刻辨識出國際評比閱讀素養，與臺灣國中生閱讀與語文教育的目標有極大落差。」。

力表現。由於閱讀理解測試（reading comprehension test）最常見的型態，即是「以篇章做為觸發素材（stimulus），並編製一系列具有數個選項的試題（multiple-choice test item）」（王振亞，2008：194），故循上述構想所進行的研究，如盧雪梅〈國中基測國文科閱讀文本暨學生表現分析〉（2011），亦先從「閱讀選擇題組」入手。本文擬將關注的焦點放在「高中、高職升大學（含科技校院）階段，從「學科能力測驗」、「指定科目考試」（以上升普通大學）、「四技統一入學測驗」（升科技校院）[3]的國文科試題中，選取「白話文閱讀選擇題組」為研究對象，除了想藉以了解國內閱讀能力檢測的部分面向，也希望透過上述三項考試的實徵資料，建構未來高中、高職階段閱讀能力檢測的發展基礎。

二、研究對象

本文將研究對象——「學科能力測驗」、「指定科目考試」、「四技統一入學測驗」（以下簡稱「學測」、「指考」、「統測」）國文科試題中的「白話文閱讀選擇題組」（以下簡稱「白話文題組」），設限於「91 至 100 學年度（以下「學年度」簡稱「年」）」的範圍內，主要是因為「指考」起始於 91 年。而「統測」雖起始於 90 年，但因該年國文科正好沒有「閱讀選擇題組」，故在此範圍內，只暫時排除了 83 至 90 年學測國文科的「白話文題組」。

用於閱讀能力檢測的選擇題組，概以一份閱讀素材為中心，環以二至數個選擇題。學測國文科與指考國文科另有非選擇式的閱讀題，例如 99 年學測國文科，考生須先閱讀王家祥〈秋日的聲音〉中一段文字：

[3] 「學科能力測驗」與「指定科目考試」是由大學入學考試中心提供，「統一入學測驗」是由技專校院入學測驗中心提供。「四技」，指的是科技大學、技術學院中的四年制學士班；技專校院入學測驗中心另提供科技大學、技術學院中的二年制學士班入學之用的「二技統一入學測驗」，因應試者不是高中、高職階段的學生，故不納入本文研究的範圍。

其實季節是萬物心境的轉換；秋日的天空時常沒有欲望，看不見一抹雲彩，秋高氣爽似乎意味著心境圓滿的狀態。春日的新生喜悅，叨叨絮絮，到夏日的豐盈旺盛，滿溢狂瀉；風雨之後，秋日是一種平和安寧的靜心，內心既無欲望也就聽不見喧囂的聲音，此時真正的聲音便容易出現了。秋天似乎是為了靜靜等待冬日的死亡肅寂做準備，曠野上行將死亡的植物時常給我們憂鬱的印象，所以誤以為秋天是憂傷的季節。也許秋天的心境讓我們容易看見深層的自己，彷彿這是天地的韻律，存在已久，只是我們習於不再察覺。

然後回答三個問題：①作者對「悲秋」傳統有何看法？②作者認為萬物的心境與四季的轉換有何相應之處？③「真正的聲音」從何而來？這類題目雖然屬於閱讀能力檢測，且亦是由數個問題組合而成的題組，但因為不是選擇題，故不納入本文的研究對象。

在學測國文科、指考國文科中，凡以「選擇題組」型態出現者，大抵為閱讀能力檢測，極少數如 93 年指考國文科的下列選擇題組，以考察古代文化知識為主，故不計入「閱讀選擇題組」：

①依據目錄推測，該書最可能在介紹：

(A)儒學思想　(B)政治思想

(C)區域文化　(D)文學批評

②下列經典，與戊、己、庚三者所討論的課題，關係最疏遠的選項是：

(A)《孟子》　(B)《荀子》

(C)《春秋》　(D)《中庸》、《大學》

甲、孟軻論「人有四端」
乙、荀況論「禮治」
丙、董仲舒論「春秋大義」
丁、韓愈的「排斥佛老」
戊、程頤論「格物窮理」
己、朱熹論「存天理，去人欲」
庚、王守仁的「致良知」

又在統測國文科中，由於 99 年起才恢復主觀式的寫作能力檢測，所以在 91 至 98 年的試題中，有少數列於「語文表達能力測驗」這個大題的「選擇題組」，其實是客觀式的寫作能力檢測，如 97 年統測國文科：

根據行政院公布的調查統計資料推估，未來十年的人口比率，五十歲以上的人口，將佔全國人口數的三分之一強。高齡長輩的居家安全，是購屋者列入重要考量的必要條件。為防止家中老人發生意外，居家內部的監控系統，將是未來「數位家庭」的標準配備。

①上文劃線處的文句，應如何修改才能清晰通順？

(A)未來十年，五十歲臺灣人口，將佔全國三分之一強

(B)未來十年，臺灣五十歲以上，將佔全國比率的三分之一強

(C)未來十年，臺灣五十歲以上的人口，將佔全國總人口的三分之一強

(D)未來十年，五十歲的臺灣人口數，將佔全國人口三分之一強的比率

②「是購屋者列入重要考量的必要條件」一句，應如何修改，才能使文句通順合理？

(A)逐漸成為購屋者的必要考量　(B)成為購屋者的必要逐漸考量

(C)逐漸考量購屋者的必要條件　(D)必要逐漸考量購屋者的條件

這個題組考察的是「文句修改」能力，雖然是「選擇題組」，但因不屬於閱讀能力檢測，也不計入「閱讀選擇題組」。

本文所以將研究的閱讀題組限於以「白話文」為閱讀素材者，係基於古代文言文、舊體韻文的教學，目的之一是讓學生熟悉古代漢語知識和古代文化知

識，因此，以之為閱讀素材的試題，也一定會出現這類知識的考察，例如：

下列關於「代朕君汝者也」與「不知乃更有儲君」句中兩個「君」字的敘述，何者正確？

(A)兩者皆為名詞　　(B)兩者皆為動詞

(C)前者為名詞，後者為動詞　　(D)前者為動詞，後者為名詞

（94 年統測國文科，閱讀素材為《資治通鑑‧唐紀三十一》）

下列關於此詩「格律、風格、文體」的敘述，何者正確？

(A)屬於五言古詩　　(B)沒有對仗句法

(C)風格浪漫飄逸　　(D)深、心、金、簪押韻

（96 年統測國文科，閱讀素材為杜甫〈春望〉）

而以「白話文」為閱讀素材的試題，通常不會對學生有這樣的閱讀要求。上述盧雪梅（2011）的研究，由於將所有國中基測閱讀題組一概納為研究對象，故原先「採 PISA 架構將試題測量的閱讀歷程分為五類」的構想，不得不因為「後來發現考詞性、語詞結構、修辭方法、錯別字、標點符號、特定的語文知識但無關閱讀理解的題目還不少」，「不屬於 PISA 的閱讀歷程」，而於「閱讀歷程編碼」進行調整。本文有鑑於此，乃暫以「白話文」閱讀素材的「選擇題組」為研究範圍。但須特別說明的是，本文為討論之便，所取「白話」題材，不全以「現代」為限，例如 97 年學測國文科某選擇題組的閱讀素材取自南宋朱熹《朱子語類》：

甲、近日學者病在好高，讀《論語》，未問「學而時習」，便說「一貫」；《孟子》，未言「梁王問利」，便說「盡心」。

乙、或問：「孟子說『仁』字，義甚分明，孔子都不曾分曉說，是如何？」

曰：「孔子未嘗不說，只是公自不會看耳。譬如今沙糖，孟子但說糖味甜耳。孔子雖不如此說，卻只將那糖與人吃。人若肯吃，則其味之甜，自不待說而知也。」

或如 93 年指考國文科某選擇題組的閱讀素材取自古代話本小說《水滸傳》，它們雖然都是古代文獻，但因接近今日語言，故亦列入白話。

經歸納整理，表 1 顯示：91 至 100 年學測、指考、統測的國文科計有 93 個「閱讀選擇題組」，考了 221 題，其中「白話文題組」有 46 個閱讀素材、108 個試題，即本文設定的研究對象。三項考試分別來看，學測國文科和統測國文科都是「白話文題組」略多於「文言文（含舊體韻文）題組」，比例大約是 1：1；指考國文科則「白話文題組」明顯少於「文言文（含舊體韻文）題組」，比例大約是 1：2。

表1　三項考試閱讀選擇題組數量表

	學測		指考		統測		總計	
	組數	題數	組數	題數	組數	題數	組數	題數
白話文題組	11	22	7	16	28	70	46	108
文言文題組	10	21	13	28	24	64	47	113
全部閱讀題組	21	43	20	44	52	134	93	221

藉由表 2，可觀察 91 至 100 年學測、指考、統測國文科對「閱讀選擇題組」的使用程度。其中統測國文科最常用「閱讀選擇題組」，佔該考科選擇題的 28.2%；學測國文科和指考國文科的「閱讀選擇題組」，則分別只佔該考科選擇題的 18.7%、17.6%。若從歷時的角度觀察，學測國文科和指考國文科在這十年中，「閱讀選擇題組」並不是年年都有的題型，但 99 和 100 年的四份題本，都維持各考 3 篇、合計 6 題（每篇 2 題）；但統測國文科在這十年中，不僅年年都有「閱讀選擇題組」，且比重一直增加——在還沒恢復考作文、全卷

50 個選擇題的前八年，已從 14%上升到 38%；99 和 100 年恢復考作文後，全卷選擇題減為 38 題，各題本的「閱讀選擇題組」仍有 18 題（考 6 篇，每篇 3 題）之多，佔全卷選擇題的 47.4%。

既然學測國文科和指考國文科在這十年中，沒有年年都考「閱讀選擇題組」，自然也無法形成考「白話文題組」的慣例。至於十年來年年都有「閱讀選擇題組」的統測國文科，以表 2 觀之，應該已形成考「白話文題組」的傳統了。

表2　閱讀題組試題數與選擇題總數對應表

	學測				指考				統測			
	題組試題數			選擇題總數	題組試題數			選擇題總數	題組試題數			選擇題總數
	白話	文言	合計		白話	文言	合計		白話	文言	合計	
091	0	0	0	24	0	8	8	24	3	4	7	50
092	0	5	5	23	2	4	6	29	2	6	8	50
093	2	4	6	22	6	2	8	29	7	3	10	50
094	4	2	6	23	0	4	4	24	6	5	11	50
095	0	0	0	23	0	0	0	24	4	9	13	50
096	2	2	4	23	0	0	0	24	9	5	14	50
097	2	2	4	23	2	2	4	24	7	9	16	50
098	4	2	6	23	0	2	2	24	11	8	19	50
099	4	2	6	23	4	2	6	24	12	6	18	38
100	4	2	6	23	2	4	6	24	9	9	18	38
總計	22	21	43	230	16	28	44	250	70	64	134	476

三、研究方法

由於「閱讀選擇題組」是針對一份閱讀素材設計二至數個試題，故本文擬

就設定為研究對象的「白話文題組」，分為「閱讀素材」與「試題」兩方面探討。對於 46 個閱讀素材，留意的是其形式與文體類別為何？對於 108 個試題，則留意它們涉及哪些閱讀層次？考生對不同閱讀層次的試題，作答表現有何不同？為了達到預期的研究目標，本文除了須憑藉相關文獻的輔助，也將整合建構一套適合分析研究對象的閱讀層次。至於考生作答表現的數據資料，學測國文科與指考國文科取自「財團法人大學入學考試中心」網站（http://www.ceec.edu.tw），統測國文科則係「財團法人技專校院入學測驗中心」惠賜，謹此致謝。

以下，本文將依「素材分類與閱讀層次」、「試題觀察與考生表現」分別討論，最後提出「結論與建議」。

貳、素材分類與閱讀層次

一、閱讀素材的分類

依據上文，91 至 100 年學測、指考、統測的國文科共考了 46 個「白話文題組」，那麼，這 46 個題組的閱讀素材是屬於何種文章？

PIRLS 將閱讀目的分為兩類：一是為了文學體驗，二是為了獲取並運用訊息（張穎，2006）。美國 NAEP[4]則將閱讀目的分為三類：一是為了文學體驗，二是為了獲取訊息，三是為了執行任務（如閱讀火車時刻表、稅率計算說明等）（孫明峰、廖純英，2006），這三類其實與 PIRLS 的兩類差不多，只是將 PIRLS 的實用型閱讀細分為二。PISA 則將閱讀目的分為四類：一是為了個人興趣，二是為了社會參與，三是為了執行工作，四是為了教育學習（樂中保，2008）。若參考中國大陸依據「普通高中語文課程標準（實驗）」（俗稱「新課標」）所制定的「高考考試大綱」，將「選考內容」分為「文學類文本閱讀」、「實

[4] 美國「全國教育進展評量測驗」（National Assessment of Educational Progress，NAEP）針對 4 年級、8 年級和 12 年級的學生，每四年進行例行評量，以追蹤其學習成就的進展趨勢。

用類文本閱讀」兩類（中國教育線上高考頻道，2011），則上述中 PIRLS 的分類，似較符合國文教師對測驗素材的看法，因此本文擬採用之，在閱讀目的上以「文學文本」與「非文學文本」來區分本文研究對象的 46 個題組閱讀素材。

此外，PISA 將閱讀素材的形式分為四種（廖先、祝新華，2010）：（一）連續性文本（continuous text）：由文句組成的段落或篇章；（二）非連續性文本（non-continuous text）：如統計圖、表格、型錄、節目單、邀請函等；（三）綜合文本（mixed text）：包含連續性文本和非連續性文本；（四）多重文本（multiple text）：數個獨立文本的組合。這對目前學測、指考、統測國文科「白話文閱讀題組」的素材選擇頗具啟發意義，因此，本文也擬用以觀察 46 個白話閱讀素材的文本形式。

二、閱讀層次的設定

（一）文獻探討

不少前輩學者認為，閱讀由淺至深可分為字面理解、解釋、評鑑三個進程：

> 許多學者用不同的研究方法對閱讀理解的層次進行探討，其中有代表性的當推 F.B.戴維斯的研究，……提出了閱讀技能的五個要素：①把握詞義，②就一定的問題，從讀物中尋找直接或間接的答案，③就讀物內容進行推論，④識別作者的意圖、態度、語氣和基調，⑤了解讀物的內部結構。……此後，不少學者（如赫伯、巴儒特、笛康和史密斯等）把閱讀技能和理解層次結合起來，把閱讀理解分成三級水平，即：①字面的理解，②解釋，③評價。這種畫分被普遍接受，而且與戴維斯的結論相吻合：戴氏的①②兩項相當於「字面理解」水平，③項相當於「解釋」水平，④⑤兩項近似於「評價」水平。（章熊，2000：314）

PISA 所提出的閱讀層次可說是與之一致。雖然 PISA 認為，閱讀在理論上有如圖 1 所示的 a、b、c、d、e 五個層次，但 PISA 在實際測試時，還是併為「入門與擷取」、「整合與解釋」、「反思與評鑑」三個層次（樂中保，2008；廖先、祝新華，2010）。

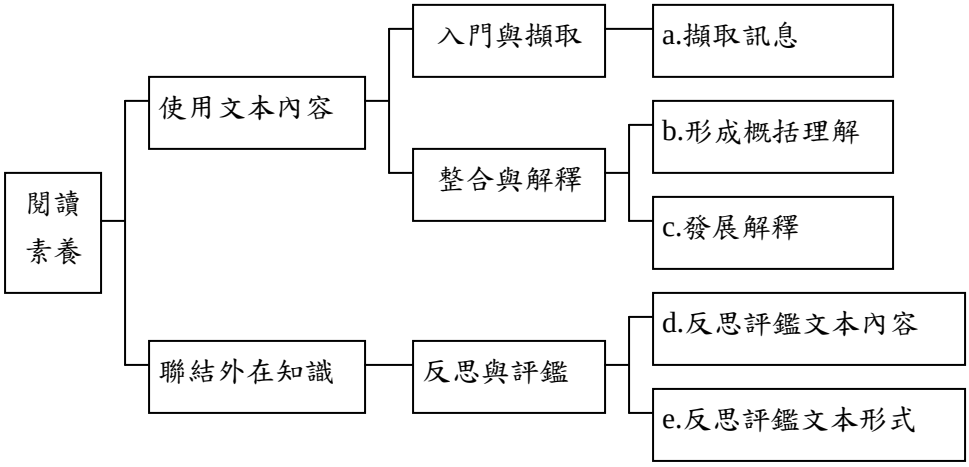


圖1 PISA閱讀層次

PIRLS 將閱讀層次分為：提取明確訊息、直接推論、解釋並整合訊息、評鑑文本的語言和內容（張穎，2006）。與 PISA 所提出的閱讀層次相較，PIRLS 的「提取明確訊息」和 PISA「入門與擷取」相同，PIRLS 的「直接推論」、「解釋並整合訊息」都可含括於 PISA 的「整合與解釋」中，PIRLS 的「評鑑文本的語言和內容」也相當於 PISA 的「反思與評鑑」——包括反思評鑑文本內容和反思評鑑文本形式。

美國 NAEP 則將閱讀層次分為：整體感知、形成解釋、聯結讀者經驗、評鑑文本內容和形式（于燕，2006）。和 PISA、PIRLS 相比，似乎特別強調「讀者」的參與。不過，一則 NAEP 是要學生根據文本訊息做為論述憑藉，並非讓學生隨意表達主觀好惡（孫明峰、廖純英，2006：82）；再則當代文學理論原就主張：作品中必有「空白」等候讀者聯想補充，故其意義實來自與讀者的交

互作用[5]；因此，NAEP 的「整體感知、形成解釋、聯結讀者經驗」，其實與 PISA 的「形成概括理解、發展解釋」、PIRLS 的「直接推論、解釋並整合訊息」一樣，都是介乎「字面理解」和「評鑑」之間的理解歷程，只是將此歷程再做若干分解時，選取的相對要素不同而已——或從「概略／細部」區分，或從「直接／隱微」區分，或從「作品／讀者」區分。

PISA、PIRLS 和 NAEP 在「評鑑」層次，都認為除了針對「文本內容」外，也會針對「文本形式」。這是因為大凡語文科目，都是在「言語內容」之外更重視「言語形式」（余應源，2001）。學習其他科目時，我們只需知道教科書的「內容」說了什麼——某數學定理的意義、某縣地理環境的特色、某重大歷史事件的始末等，但學習國文、英文等語文科目，則除了知道文章「說什麼」之外，還會繼續探討文章「怎麼說」——用了哪些詞彙才這麼生動、用了什麼句型才這麼有力、段落是怎樣布局才這麼嚴謹……，亦即追究：文章的各個局部，是如何扣組在一起而構成一個整體的。要進行這樣的閱讀，PISA 認為必須「聯結外在知識」才能做到——畢竟「操千曲而後曉聲，觀千劍而後識器」，要指出寫作方式的精良與否，必得透過其他文章的比較與閱讀經驗的累積。

NAEP 刻意將「聯結讀者經驗」區別於「整體感知、形成解釋」，目的應是想指出：閱讀有在文本範圍內的，也有須跨出文本以外的。雖然誠如上述，當代文學理論已經否認文本的意義是「自給自足」，亦即不可能有只限於文本內的閱讀，但若檢視 91 至 100 年學測、指考、統測國文科的「白話文閱讀題組」，還真有少數試題是特地要考「延伸至文本外」的知識——這當然也是閱讀文章的一種理解，但與 PISA、NAEP 仍「集中於文本內」、只是提示要加入個人經驗的理解，頗不相同。

---

[5] 龍協濤《讀者反應理論》：「文學作品的意義並不像傳統詮釋學所認為的那樣是作品中原先固有的。……作品的意義只有在閱讀過程中才能產生。……作品的未定性與意義空白，促使讀者去尋找作品的意義，它呼籲、導引、召喚、邀請讀者，賦予他參與作品意義構成的權利。因此，未定性與意義空白構成作品的基礎結構，這就是伊舍（Wolfgang Iser）的『召喚結構』。……正是在一部作品意味深長的沉默中，在它的意義空白中，隱藏著作品效果的潛能。」（頁 101-104）

（二）本文的設定

本文基於以上述評，將以下列整合而成的閱讀層次，繼續探討研究對象。這個閱讀層次仍先區別出「集中於文本內」、「延伸至文本外」，說明如下。

1、集中於文本內：四個閱讀層次

（A）尋找局部和表層訊息

在這個閱讀層次，試題不會太讓人費神思索，讀者只要依據題幹的要求，到文章中去搜尋明顯出現於文章表面的訊息。有些訊息可能是跨段落的，但讀者在確認的過程中，並不需要理解全文的涵義。例如：

> 依據上文，糖精與阿斯巴甜的共同特點是：
> (A)可耐高溫　　(B)不會產生熱量
> (C)甜度是蔗糖的三百倍　　(D)糖尿病患者可以食用
> （93 年統測國文科，閱讀素材改寫自郝龍斌《健康飲食 Go, Go, Go！》）

> 依據上文，下列關於「伽利略」的敘述，何者正確？
> (A)師承哥白尼，提出「日心說」
> (B)成為宮廷數學家之後，其專業權威始獲肯定
> (C)發現太陽系中的木星，取名為「麥第奇之星」
> (D)用金錢賄賂麥第奇公爵，換取地位較高的工作
> （98 年統測國文科，閱讀素材改寫自英家銘、蘇意雯〈數學與「禮物交換」〉）

上引兩題的閱讀素材，一篇談述數種「代糖」，「糖精」與「阿斯巴甜」只是其中兩種；一篇則說明科學研究的進展、專家地位的形成，有其政治社會因素，「伽利略」只是引述的例子。讀者回答這兩題，均可「但照隅隙」，暫時免觀

全文。不過有時候，讀者要確認的訊息，也可能沒那麼直接而明顯。例如：

依據上文第三段的敘述，下列哪一種農作物可不需藉由蜜蜂授粉？

(A)稻子　(B)大豆　(C)棉花　(D)蘋果

（99 年統測國文科，閱讀素材改寫自 Alison Benjamin、Brian McCallum《蜜蜂消失後的世界》）

讀者須從文章中的片段：「早餐只有一成不變的清粥白飯、麥片，沒有新鮮果汁，也沒有豆漿」、「因為棉花產量大減，棉製衣服將貴得不像話」，推知「清粥白飯」是用「稻子」收成後的米所煮成，才能確定答案。

（B）形成統括和概約理解

在這個閱讀層次，讀者已經通讀全文，所以有了「統括」這個寬窄度上的認知，但在深淺度上可以是「概約」，屬於縱觀瀏覽後對文章的初步了解，可能包括文章的主題、所述事件的大致經過、故事的主要背景和人物等，例如：

依據上文，自 1980 年代中期至神經病理學家蒙特這段期間，關於胰島素的科學研究進程是：

甲、發現大腦會分泌胰島素

乙、發現糖尿病導因於胰島素分泌異常

丙、發現阿茲海默症患者的大腦胰島素含量低

丁、發現記憶力好壞與大腦胰島素分泌多寡有關

(A)甲→乙→丁　(B)甲→丁→丙 (C)乙→甲→丁 (D)乙→甲→丙

（98 年學測國文科，閱讀素材改寫自 Melinda Wenner 著，林雅玲譯，〈大腦也會得糖尿病〉）

金融市場常言：「行情總在絕望中誕生，在半信半疑中成長，在憧憬中茁壯，在充滿希望中破滅」。上文敘述的鬱金香價格漲跌經過，沒有上述行情演變的哪一個階段？

(A)在絕望中誕生　(B)在半信半疑中成長

(C)在憧憬中茁壯　(D)在充滿希望中破滅

（98 年統測國文科，閱讀素材改寫自墨基爾著，楊美齡、林麗冠譯，《漫步華爾街‧鬱金香狂熱》）

而有些試題，雖然題幹問的是文章中某一詞語，但該詞語原來就是全文主題，且其意義只需略讀全文就能直接推得，則仍應屬「形成統括和概約理解」而非「發展細部和深層解釋」，例如：

詩中的「你」是指：

(A)想念的人　(B)臺灣　(C)美人　(D)太平洋（92 年統測國文科）

這題的閱讀素材是李魁賢〈島嶼臺灣〉：「從空中鳥瞰／被你呈現肌理的美吸引／急切降落到你的身上／你是太平洋上的／美人魚／我永恆故鄉的座標」，從詩題即可推知「你」是代稱「臺灣」。或如下題：

詩中「一彎弧線」指的是：

(A)唇形　(B)眉形　(C)眼形　(D)臉形（91 年統測國文科）

這題的閱讀素材是林廣〈微笑〉：「一彎弧線，輕柔地／承載天地所有的重負／一條會唱歌的溪流／把一粒一粒頑石／雕成剔透的水晶」，答案也很明顯就是詩題。

（C）發展細部和深層解釋

這個層次是在「概約」程度的理解上，繼續掘發文章的涵義。此時，文章的意義既來自「自下而上的加工」——從詞到句再到篇循序完足，也來自「自

上而下的加工」[6]——即上文所述：讀者會主動填補作品中的「意義空白」。理解於是展開「詮釋循環」（hermeneutic circle），這包括了：（1）理解總是整體的理解，任何一部分理解的變動都會改變整體的理解，（2）只有在共同的思想方式及文化背景上，理解活動才可能進行，（3）意義不是由作品流向解釋者，而是解釋者與作品在不斷變換整體與部分地位的關係中，相互交流而形成[7]。也正因為在「細部和深層的解釋」過程中，「整體」與「細節」之間是因彼尋覓此、因此認取彼，「整體」與「細節」的意義也是互相填補、互相註解才形成，所以，儘管讀者在這個層次必須使出分類、比較、歸納、推論等能力，但既已交錯互用，也就不必強行割裂。

這個層次的試題，小則聚焦於篇內某一詞句，大則觀照全篇，例如 92 年指考國文科以許常德〈公寓〉為閱讀素材的一題，便將大小熔於一爐：

> 詩以「海」→「河」→「井」層遞的方式為喻，頗富深意，下列的理解，何者不正確？
>
> (A)「海洋瘦成河」比喻人們的活動空間因為公寓林立而變得狹窄
>
> (B)「河無異於井」比喻居住空間的壓縮，對人們的胸懷視野造成負面影響
>
> (C)詩人認為人們原本應該「海闊天空」，可惜公寓建築使人們「坐井觀天」
>
> (D)公寓「河海不擇細流」的特性，使「離鄉背井」的外地人得以在都市棲身

比較需要說明的是，我們不能一見題幹問「全篇主旨」，便認為是「形成統括

---

[6] 閱讀的「自下而上模型」（bottom-up model）認為，閱讀是從字母到詞、到句、再到篇的有序且自下而上的歷程；「自上而下模型」（top-down model）則主張，閱讀的起點是讀者的語言認知和相關經驗，閱讀是基於讀者前知識、受概念所驅使的自上而下的歷程；「相互作用模型」（interactive model）則認為，閱讀既有自下而上的加工，也有自上而下的加工，這兩種加工同時存在於字母、詞、句、篇章層次，並交互發生作用。參閱范琳、周紅、劉振前，《二語語篇閱讀推理的心理學研究》（北京：北京大學出版社，2011 年），頁 15-17。

[7] 關於「詮釋循環」，可參考殷鼎，《理解的命運》（臺北：東大圖書公司，1990 年)，頁 33-37。

和概約理解」。關於「本篇在說什麼」的答案，文學作品往往無法「望文生義」，而必須在諸多線索暗示、意象隱喻之間反覆推敲，例如 96 統測國文科中的一題：

下列關於此詩主旨的敘述，何者正確？

(A)抒發悲天憫人的胸懷　　(B)呈顯寂寞空虛的心靈

(C)表現自信狂傲的性格　　(D)展露少不經事的稚氣

讀者閱讀的詩是紀弦〈狼之獨步〉：「我乃曠野裡獨來獨往的一匹狼／不是先知，沒有半個字的嘆息／而恆以數聲悽厲已極之長嗥／搖撼彼空無一物之天地／使天地戰慄如同發了瘧疾／並刮起涼風颯颯的，颯颯颯颯的／這就是一種過癮」。詩人自比為「獨來獨往的狼」想表達什麼，絕不可能在大略掃瞄後的概約理解這層中得知。本題有 49%的考生誤選「B」，很有可能就是憑著「獨來獨往」等於「寂寞」、「悽厲長嗥」代表「空虛」的浮泛印象進行理解。

此外，這個層次的理解也有可能是跨學科的。例如 93 年統測國文科第 32 題，要考生依據文中所述：「『十七年蟬』的生命週期所以是十七，一是因為這個數字長過潛在敵人的壽命，再則因為『十七』是個質數（除了一及本身外，沒有數字能整除它）。若天敵的生命週期是五年，則十七年蟬就能每五乘十七——即八十五年才倒楣一次。」推斷生存於美國南部、與「十七年蟬」採取相同避敵策略的「X 年蟬」，其「X」最可能是十二、十三、十六或二十。考生要答對這題，必須對「除了一及本身外，沒有數字能整除它」這個數學概念是熟悉的。

（D）掌握敘寫和組織方式

不同於上述（A）、（B）、（C）均屬「文本內容」的理解，這個層次是對「文本形式」的理解。PISA 將這個層次稱為「反思評鑑文本形式」。由於

國內常以 Bloom 的「認知領域教育目標分類」中對「評鑑」向度的定義[8]來了解「評鑑」一詞，但從〈臺灣 PISA2009 結果報告〉所提供的樣本試題來看（臺灣 PISA 國家研究中心，2011：76），屬此層次的試題[9]比較接近 Bloom「認知領域教育目標分類」中的「分析」向度[10]，故本文不特別強調此一對「文本形式」的理解是不是「評鑑」，尤其是透過選擇題的方式檢測時，即使試題真的對閱讀素材進行「評鑑」，主要也是命題者而非讀者的察覺和判斷。

對「文本內容」的理解，是關切文章「說什麼」；對「文本形式」的理解，則注意的是文章「怎麼說」，即掌握文章的各個局部是如何扣組在一起而構成一個整體。例如：

關於上文的敘寫線索，正確的選項是：
(A)先寫九份的山巒，再寫九份的海濱
(B)先寫九份的黃昏，再寫九份的夜色
(C)先寫九份廟宇殘破，再寫九份居民蒼老
(D)先寫九份昔日的盛況，再寫九份今日的沒落
（97 年指考國文科，閱讀素材為古蒙仁〈破碎了的淘金夢〉）

上文述及《樂舞圖》和《韓熙載夜宴圖》，目的是為了說明：
(A)椅子出現於古代中國，首見於唐代
(B)唐代椅子尚未普及，跪坐仍是主流坐姿
(C)唐代中期以後，垂足坐在椅子上成為主流坐姿
(D)唐代平時多採跪坐，宴會則喜歡垂足坐在椅子上
（100 年統測國文科，閱讀素材改寫自濬臺卓爾《椅子改變中國》）

---

[8] 「評鑑」（evaluate）是「根據標準下判斷」，可分為「檢查」（checking）：「檢視某程序或產品中的不一致性或錯誤，確定某程序或產品的內部一致性，察覺正實行程序的效能」，「評論」（critiquing）：「檢視產品和外部規準的不一致性，確定產品是否有外部一致性，察覺解決問題的方式適切性」。參閱葉連祺、林淑萍，〈布魯姆認知領域教育目標分類修訂版之探討〉，《教育研究月刊》105 期（2003 年 1 月）。

[9] 〈臺灣 PISA2009 結果報告〉中只有一個「反思評鑑文本形式」的樣本試題，為單一選擇題：「為什麼文中提到筆？(A)幫助你理解怎樣握牙刷；(B)因為你同時以筆和牙刷從一側開始；(C)表明你可以用許多不同的方法刷牙；(D)因為你刷牙應像寫字那麼認真。」

[10] 「分析」（analyze）是「分解整體為許多部分，並決定各部分彼此和與整體結構或目的的關係」，又分為「辨別」（differentiating）、「組織」（rganizing）、「歸因」（attributing）。

敘寫手法也可以是跨文本進行比較。例如 93 年指考國文科提供兩段不同版本的《水滸傳》文字——〔甲〕和〔乙〕內容相同，但敘事方式不太一樣，如〔甲〕是寫：「武松也把眼來虛閉緊了，撲地仰倒在凳邊，那婦人笑道……，只見裡面跳出兩個蠢漢來……」，〔乙〕則是寫：「武松也雙眼緊閉，撲地仰倒在凳邊，只聽得笑道……，只聽得飛奔出兩個蠢漢來……」，然後讓考生回答下題，分辨兩種寫法的差異：

若〔乙〕是〔甲〕的修改版，〔乙〕的改動主要在：

(A)刪除冗贅文句，減輕閱讀負荷

(B)讓孫二娘由謀財者變害命者，擴大情節起伏

(C)增添「老娘」、「鳥男女」、「鳥大漢」等詞，以襯托孫二娘的粗鄙與兇狠

(D)將武松對週遭的了解，改為由「聽」、「想」而得，以符合雙眼緊閉的狀態

2、延伸至文本外：一個閱讀層次

（E）聯結相涉的語文知識

上文已言及，PISA 和美國 NAEP 在編擬閱讀層次時，雖然指出了「聯結外在知識」、「聯結讀者經驗」，但那只是對「理解來自文本與讀者交互作用」的提示，要讀者回答的問題仍「集中於文本內」。但學測、指考、統測的國文科，是依據特定課程大綱所編製的考試，基於課程學習內容包含了特定的語言和文化知識，因此會在「白話文題組」中，偶爾見到刻意「延伸至文本外」、要考生回顧那些知識的測試。但不能否認的是，這也是對文章的一種理解，甚至是涉及文化背景而層次更深的理解。例如下列截圖來自某品牌潤喉糖的電視廣告，讀者知道、或不知道它運用「孟姜女哭倒長城」的典故，體會是絕對不同的。

圖2　某品牌潤喉糖電視廣告截圖

在這個閱讀層次，那些因閱讀文章而可能涉及的語言和文化知識，是被試題特意凸顯出來的，也就是說，將這些知識運用於理解過程，對所閱讀的文章而言不是絕對必然，而是別開「深」面的觸類引申。例如 91 年統測國文科，試題於林廣〈微笑〉詩中找出一句，延伸去考修辭格裡的「擬人」手法：

「一條會唱歌的溪流」運用「擬人」的寫作手法，下列流行歌詞，何者也有相同的表現手法？

(A)海的那一邊，烏雲一整片，我很想為了你快樂一點

(B)陪你去看流星雨落在這地球上，讓你的淚落在我肩膀

(C)廣場一枚銅幣，悲傷得很隱密，它在許願池裡輕輕歎息

(D)半夜睡不著覺，把心情哼成歌，只好到屋頂找另一個夢境

閱讀這首詩時，未必要注意「一條會唱歌的溪流」是用了哪一種修辭格，大概也不會再去聯想到其他使用「擬人」的流行歌詞。又如下列兩題，則延伸到課程中一些重要文化經典的學習內容：

「像一隻老青蛙的青蛙」的言論，最接近先秦思想的哪一家？

(A)儒家　　(B)道家　　(C)法家　　(D)縱橫家

（93 年統測國文科，閱讀素材節選自芥川龍之芥〈青蛙〉）

依據上文所敘述的坐姿演變，下列文句的「坐」，按句中人物所處時代推斷，坐姿屬於「跪坐」的是：

甲、子路、曾晳、冉有、公西華侍「坐」

乙、項王即日因留沛公與飲，項王、項伯東嚮「坐」

丙、劉老老入了「坐」，拿起箸來，沉甸甸的不伏手

丁、魯肅領了周瑜言語，逕來舟中相探孔明，孔明接入小舟對「坐」

戊、三個來到酒店裡，宋江上首「坐」了；武松倚了哨棒，下席坐了

(A)甲乙丙　(B)甲乙丁　(C)乙丁戊　(D)丙丁戊

（100 年統測國文科，閱讀素材改寫自濬臺卓爾《椅子改變中國》）

要回答這兩題，讀者必須了解「先秦時期」的「儒家」、「道家」、「法家」、「縱橫家」有什麼特色，必須熟悉「子路、曾晳、冉有、公西華」、「項王、沛公」、「劉老老」、「周瑜、孔明」、「宋江、武松」等各組人物是哪些書中所提到的什麼時代的人。這些同樣是在一般閱讀文章時不會特別想到，且沒想到也不至於影響基本理解。

雖然讀者會因這類「延伸至文本外」的試題而暫時無法「集中於文本內」，但事實上，讀者卻也會被這些乍看之下「偏離主題」的試題所引導，轉去回想那些潛藏於記憶中的語言知識、文化知識，而正是這些知識所累積的基礎，才能應付各種閱讀任務。所以，這些看似歧出的理解，雖屬通常未必會想到、即使沒想到也無甚影響的理解，但透過命題者的指點，未嘗不是一條可以啟動記憶中的語言和文化知識、因而可以看到更多人文景觀的閱讀路徑。

參、試題觀察與考生表現

一、閱讀素材的類別

表 3 顯示：91 至 100 年學測、指考、統測國文科「白話文題組」的 46 個

閱讀素材，幾乎全屬 PISA 素材分類中的「連續性文本」，也就是考「單一的」文字篇章，佔 85%；另有三個「多重文本」其實也是文字篇章，只是因在同一題組中提供兩篇（段）出處不同、或出自同書但不相連的文字[11]，所以才分別計算。

「綜合文本」是「文字加圖表」的閱讀素材，在 46 個閱讀素材中只出現過四次：96 年統測國文科 38-40 題是一文一圖，取材自 Rita Carter 著、洪蘭譯的《大腦的秘密檔案》第六章；98 年統測國文科 19-20 題是一文一表，取材自簡媜《老師的十二樣見面禮》；99 年學測國文科 14-15 題是一文一圖，取材自森本哲郎《一個通商國家的興亡》；99 年統測國文科 24-26 題也是一文一圖，取材自 Alfred W. Crosby《哥倫布大交換》[12]。至於「非連續性文本」，則在 46 個閱讀素材中從未出現過。

整體來看，這三項升學考試的國文科都極少使用圖、表等其他非文字篇章做為「白話文題組」的閱讀素材，這應該與高中、高職國文課程向來的教學內容有關。

表3　白話文閱讀題組文本類型表

	學測	指考	統測	總計
連續性文本	9	6	24	39
非連續性文本	0	0	0	0
綜合文本	1	0	3	4
多重文本	1	1	1	3

表 4 顯示：91 至 100 年學測、指考、統測國文科「白話文題組」的 46 個

[11] 三個多重文本題組為：（一）93 指考國文科 15-17 題，閱讀素材是兩段內容相同、但版本不同的《水滸傳》；（二）97 學測國文科 14-15 題，閱讀素材是《朱子語類》兩則；（三）97 統測國文科 25-26 題，閱讀素材是「主計處統計資料」和〈邊緣鬥士・愛爾蘭〉並列。

[12] 99 年學測國文科和 99 年統測國文科這兩個題組所附的地圖，雖然不看也不至於無法回答試題，但因對文章意義的理解仍有輔助之效，故仍計入「綜合文本」。

閱讀素材，有 28 個屬於「文學文本」（配置 62 題），18 個屬於「非文學文本」（配置 46 題），比例大約是 6：4，「文學文本」還是比較容易成為測驗素材。

若將三項考試的「白話文題組」分開來看，指考國文科幾乎都考「文學文本」（7 篇中佔 6 篇），學測國文科的「文學文本」也超過七成（11 篇中佔 8 篇），只有統測國文科是「文學文本」與「非文學文本」各佔一半（各 14 篇），甚至以題數來計算，「非文學文本」所考的題數還多一些。若綜觀三項考試的十年歷程，也只有統測國文科是自 93 年以後，便在「白話文題組」形成考「非文學文本」的傳統。至於在學測國文科與指考國文科方面，雖然考「白話文題組」在近四、五年漸成常態，但還沒給「非文學文本」固定的席次。

表4　文學文本題組與非文學文本題組數量表

	學測		指考		統測		總計	
	組數	題數	組數	題數	組數	題數	組數	題數
文學文本題組	8	16	6	13	14	33	28	62
非文學文本題組	3	6	1	3	14	37	18	46
白話文題組	11	22	7	16	28	70	46	108

二、試題所屬的閱讀層次

表 5 顯示：91 至 100 年學測、指考、統測國文科「白話文題組」的 108 個試題中，有近乎一半（49.07%）屬於「發展細部和深層解釋」，其次依序是：「形成統括和概約理解」（24.07%）、「尋找局部和表層訊息」（14.81%）、「掌握敘寫和組織方式」（7.41%）、「聯結相涉的語文知識」（4.63%）。也就是說，差不多 88%的試題都集中於「文本內容」的理解，而這也充分反映出我們通常對閱讀文章的想法——以把握訊息為優先，最主要是知道「說什麼」；行有餘力，則看一下「怎麼說」；至於「延伸至文本外」，就看是否靈光一閃了。

若將三項考試的「白話文題組」分開來看，學測國文科和指考國文科在命

題考量上，都以「發展細部和深層解釋」（學測國文科佔 22 題中的 63.6%，指考國文科佔 16 題中的 50%）、「形成統括和概約理解」（學測國文科佔 22 題中的 31.8%，指考國文科佔 16 題中的 25%）居前兩位。反觀統測國文科，雖然也以「發展細部和深層解釋」考得最多（佔 70 題中的 44.3%），但「尋找局部和表層訊息」卻能與「形成統括和概約理解」並列居次（均佔 70 題中的 21.4%），其原因自是與統測國文科考比較多「非文學文本」有關。

如表 6 所示，「文學文本題組」和「非文學文本題組」所著眼的閱讀層次頗有差異——就「文學文本題組」而言，有 62.9%的試題置於「發展細部和深層解釋」，對「形成統括和概約理解」（16.13%）和「掌握敘寫和組織方式」（11.29%）也較注重，「尋找局部和表層訊息」則顯得無關緊要（僅 3.23%）。但在「非文學文本題組」中，則較忽略「文本形式」層次（僅 2.17%），且對「發展細部和深層解釋」（30.43%）、「形成統括和概約理解」（34.78%）、「尋找局部和表層訊息」（30.43%）這三個「文本內容」層次都不偏廢。足見對於不同類型的文本，預期的閱讀層次也不一樣。

表5　各閱讀層次試題數量表

	學測題數	指考題數	統測題數	合計	在 108 題所佔比重
尋找局部和表層訊息	0	1	15	16	14.81%
形成統括和概約理解	7	4	15	26	24.07%
發展細部和深層解釋	14	8	31	53	49.07%
掌握敘寫和組織方式	0	3	5	8	7.41%
聯結相涉的語文知識	1	0	4	5	4.63%

表6　兩類文本試題所屬閱讀層次表

	文學文本題組		非文學文本題組	
	題數	比重	題數	比重
尋找局部和表層訊息	2	3.23%	14	30.43%
形成統括和概約理解	10	16.13%	16	34.78%
發展細部和深層解釋	39	62.90%	14	30.43%
掌握敘寫和組織方式	7	11.29%	1	2.17%
聯結相涉的語文知識	4	6.45%	1	2.17%

那麼，學測、指考、統測的國文科，對「文學文本題組」會不會在閱讀層次上期待不同？從表 7 看來，三項考試的「文學文本題組」都絕對以「發展細部和深層解釋」為優先。而統測國文科看來似乎對五個閱讀層次都能顧及，應該是因為統測國文科往往一個題組包含三題，相較於學測國文科和指考國文科是一個題組包含兩題，觸角自然會多一些。

但對「非文學文本題組」，三項考試的考法似乎就不太一樣了，從表 8 可見：學測國文科和指考國文科仍較偏重「形成統括和概約理解」和「發展細部和深層解釋」，統測國文科則「尋找局部和表層訊息」、「發展細部和深層解釋」、「形成統括和概約理解」三者兼顧，甚至在「尋找局部和表層訊息」上，比例還高出一點。不過，由於學測國文科和指考國文科的「非文學文本題組」很少（兩科合計 4 組，統測國文科則有 14 組），上述的比較未必概全。

表7　文學文本試題所屬閱讀層次表

	學測文學題組		指考文學題組		統測文學題組	
	題數	比重	題數	比重	題數	比重
尋找局部和表層訊息	0	0.00%	1	7.69%	1	3.03%
形成統括和概約理解	4	25.00%	2	15.38%	4	12.12%
發展細部和深層解釋	11	68.75%	7	53.85%	21	63.64%
掌握敘寫和組織方式	0	0.00%	3	23.08%	4	12.12%
聯結相涉的語文知識	1	6.25%	0	0.00%	3	9.09%

表8　非文學文本試題所屬閱讀層次表

	學測非文學題組		指考非文學題組		統測非文學題組	
	題數	比重	題數	比重	題數	比重
尋找局部和表層訊息	0	0.00%	0	0.00%	14	37.84%
形成統括和概約理解	3	50.00%	2	66.67%	11	29.73%
發展細部和深層解釋	3	50.00%	1	33.33%	10	27.03%
掌握敘寫和組織方式	0	0.00%	0	0.00%	1	2.70%
聯結相涉的語文知識	0	0.00%	0	0.00%	1	2.70%

三、考生的作答概況

學測、指考、統測國文科於「試後」所提供的難度值 P，均為全體考生在每一試題正確選項的百分比通過率，因此，雖然各年、各科的考生都不同，但因難易度 P 值的計算方式一樣，仍可在某種程度上進行跨年、跨科的歸納與比較。

另須說明的是，91 至 100 年三項考試「白話文題組」原有 108 題，但因 91 年統測國文科有一題的正答有疑義（該題屬「發展細部和深層解釋」，閱讀素材是林廣〈微笑〉，為「文學文本」），故該題無 P 值，以下表 9 和表 10 均只計 107 題的資料。

表 9 顯示：從全體考生的作答結果來看，「文學文本題組」的平均答對率（0.70）是比「非文學文本題組」的平均答對率（0.74）來得低，亦即考生較容易理解「非文學文本」，較不易理解「文學文本」，符合一般的預測。

若將三項考試的考生作答結果分別觀之，則統測國文科的考生無論在「文學文本題組」或「非文學文本題組」的理解，都較不理想。由於學測和指考是以高中生為主的升學考試，統測是以高職生為主的升學考試，倘若統測國文科的試題並非比學測和指考國文科的試題難，則可能意謂：高中生族群的閱讀理解能力約略高於高職生族群的閱讀理解能力。

表9　兩類文本試題平均答對率表

	學測		指考		統測		總計	
	題數	平均答對率	題數	平均答對率	題數	平均答對率	題數	平均答對率
文學文本題組	16	0.74	13	0.77	32	0.66	61	0.70
非文學文本題組	6	0.84	3	0.90	37	0.71	46	0.74

※統測「文學文本題組」只計 32 題，「文學文本題組」總數也只計 61 題。

表 10 顯示：分布於五個閱讀層次的 91 至 100 年學測、指考、統測國文科「白話文題組」的 107 個試題，其中屬「文本內容」閱讀層次者，考生的作答表現較佳；屬「文本形式／掌握敘寫和組織方式）」者，平均答對率便落至 0.68；屬「延伸至文本外／聯結相涉的語文知識」者，平均答對率僅 0.64。而在「文本內容」的理解上，又以「尋找局部和表層訊息」的表現最佳（平均答對率 0.78），「形成統括和概約理解」次之（平均答對率 0.73），「發展細部和深層解釋」又次之（平均答對率 0.71）。上述結果，亦符合一般的預測。

若選擇「形成統括和概約理解」和「發展細部和深層解釋」這兩個閱讀層次，分別觀察三項考試的考生作答結果，則在「形成統括和概約理解」方面：指考國文科的平均答對率是 0.86，學測國文科的平均答對率是 0.72，統測國文科的平均答對率是 0.70；而在「發展細部和深層解釋」方面：指考國文科的平均答對率是 0.80，學測國文科的平均答對率是 0.79，統測國文科的平均答對率是 0.64。如果統測國文科的試題並非比學測和指考國文科的試題難，則可能意謂：高中生族群在「形成統括和概約理解」和「發展細部和深層解釋」上，是約略高於高職生族群的。

表10　各閱讀層次試題平均答對率表

	三項考試題數	平均答對率
尋找局部和表層訊息	16	0.78
形成統括和概約理解	26	0.73
發展細部和深層解釋	52	0.71
掌握敘寫和組織方式	8	0.68
聯結相涉的語文知識	5	0.64

※「細部深層解釋」只計 52 題。

再者，綜觀 91 至 100 年學測、指考、統測國文科「白話文題組」中有 P 值的 107 個試題，當中 P＜0.50（難度偏高）者有 11 題。如表 11 所示，這 11 題繫於 9 個閱讀素材，當中 7 個都是「文學文本」。又這 11 題所屬的閱讀層次，9 題是「發展細部和深層解釋」，「形成統括和概約理解」和「聯結相涉的語文知識」則各有 1 題。據此可以推知，要就「文學文本」進行「細部和深層解釋」，對一般考生來說是較具挑戰性的。

表11　高難度試題所屬文本類型與閱讀層次表

試卷與題號	答對率	閱讀素材	文本類型	閱讀層次
94 年學測 15	0.40	夐虹〈記得〉	文學	細部深層解釋
100 年學測 13	0.38	曾貴海〈河流終將成為記憶〉	文學	統括概約理解
93 年統測 36	0.42	芥川龍之芥〈青蛙〉	文學	細部深層解釋
93 年統測 37	0.44			相涉語文知識
94 年統測 33	0.47	張曉風〈許士林的獨白〉	文學	細部深層解釋
96 年統測 28	0.39	梁啟超《飲冰室文集・雜著・笑林》	文學	細部深層解釋
96 年統測 34	0.43	紀弦〈狼之獨步〉	文學	細部深層解釋
96 年統測 35	0.43			細部深層解釋
98 年統測 22	0.42	英家銘、蘇意雯〈數學與「禮物交換」〉	非文學	細部深層解釋
99 年統測 32	0.47	錢鍾書〈窗〉	文學	細部深層解釋
100 年統測 27	0.40	澹臺卓爾《椅子改變中國》	非文學	細部深層解釋

「文學文本」比較難讀，是否就因此具有較高的鑑別度而比「非文學文本」更適合做為閱讀測驗素材呢？針對這項假設，本文從 91 至 100 年學測、指考、統測國文科「白話文題組」中，各自挑出鑑別度高居前三分之一的試題略做觀察。由於學測國文科和指考國文科是以各題「高分組答對率」減去「低分組答對率」所得的差做為鑑別度 D 值，統測國文科是以各題正確選項的點二系列相關係數做為鑑別度 R 值，D、R 算法不同，故不相互比較。

結果發現：學測國文科「白話文題組」共 22 題，鑑別度最高的前 7 題，D 值在 0.51～0.33 之間，5 題屬「文學文本題組」，2 題屬「非文學文本題組」。指考國文科「白話文題組」共 16 題，鑑別度最高的前 5 題，D 值在 0.45～0.31 之間，皆屬「文學文本題組」。統測國文科「白話文題組」共 70 題，鑑別度最高的前 21 題，R 值在 0.59～0.44 之間，11 題屬「文學文本題組」，10 題屬「非文學文本題組」。

回顧上述表 4，併計學測和指考國文科的「文學文本題組」，原就佔「白話文題組」的 76%；統測國文科的「文學文本題組」，則佔「白話文題組」的 47%。故上述高鑑別度的試題，容或稍偏集於「文學文本題組」，但情況並不明顯，尤其在「文學文本題組」與「非文學文本題組」是 1：1 的統測國文科，上述 21 個高鑑別度試題也以 1：1 分屬兩類文本，足見「考文學文本才有鑑別度」是一種成見，「非文學文本」其實也是很好的閱讀測驗素材。

肆、結論與建議

一、結論

本文試圖從「高中、高職升大學（含科技校院）」的大型入學測驗國文科試題，了解國內閱讀能力檢測的部分現況。為集中觀察視野，本文選取 91 至 100 年學測、指考、統測國文科「白話文題組」的 46 個閱讀素材、108 個試題

為研究範圍，所得到的研究結果大致如下，可做為閱讀教學與試題編製的參考：

（一）閱讀素材以連續性文本、文學文本居多

1、若併觀三項考試的國文科「白話文題組」，閱讀素材明顯集中於 PISA 素材分類中的「連續性文本」（佔 85%），也非常偏好文字篇章（佔 91%），「文字加圖表」的閱讀素材只出現過 4 次。

2、整體來說，三項考試國文科「白話文題組」在閱讀素材的選擇上，仍比較偏好「文學文本」——46 個閱讀素材中有 28 個屬於「文學文本」（配置 62 題），18 個屬於「非文學文本」（配置 46 題），題數比例大致如圖 3。

3、若將三項考試國文科「白話文題組」分別觀之，對「文學文本」的偏好程度依序是：指考國文科＞學測國文科＞統測國文科。圖 4 是以題數所計算的比例，學測和指考國文科的「文學文本題組」都佔七成以上，只有統測國文科是「文學文本題組」與「非文學文本題組」各約佔一半。

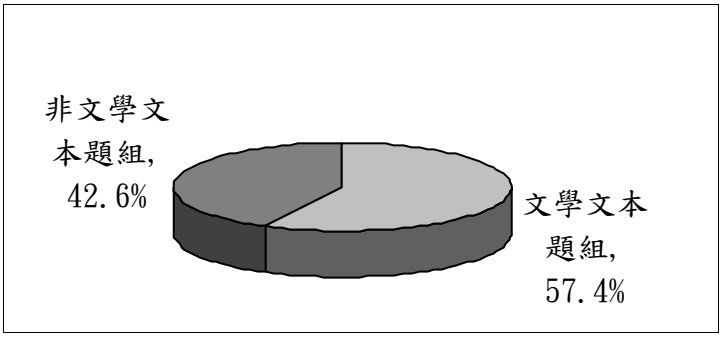


圖3 兩類文本題組的試題數量比

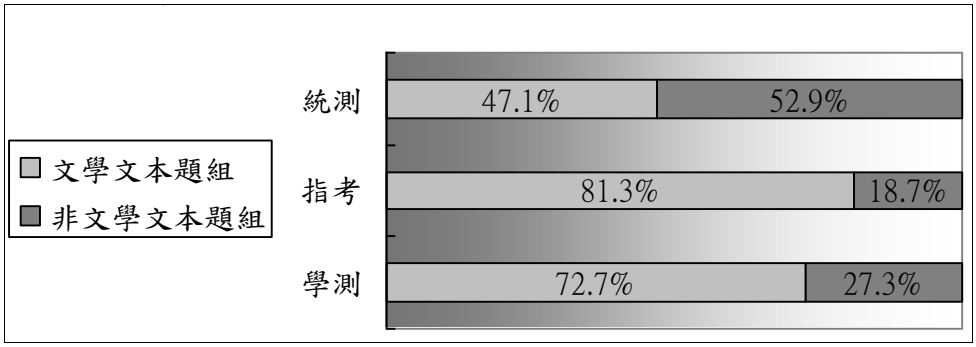


圖4 三項考試採用兩類文本的比較

（二）近半試題屬「發展細部和深層解釋」；文學和非文學文本注重不同閱讀層次

1、併觀三項考試的國文科「白話文題組」，88%的試題都集中於「文本內容」的理解，只有 7.4%的試題引導考生注意「文本形式」上的「敘寫和組織方式」，另有 6.3%的試題「延伸至文本外」，讓考生回顧記憶中儲存的「相涉語文知識」。

2、併觀三項考試的國文科「白話文題組」，如圖 5 所示：有近乎一半（49.07%）的試題屬於「發展細部和深層解釋」，有近四分之一（24.07%）的試題屬於「形成統括和概約理解」，顯示這兩個層次絕對是文章理解的重心，同時亦可見：一個題組在題數有限的情況下，能觸及較深層次理解的試題，往往是優先配置的。

3、若併觀三項考試的國文科「白話文題組」，從圖 6 和圖 7 的比較可知：「文學文本題組」比「非文學文本題組」注重「文本形式」的探究——「掌握敘寫和組織方式」的試題佔「文學文本題組」的 11.29%，但只佔「非文學文本題組」的 2.17%。又「文學文本題組」明顯偏重「發展細部和深層解釋」，有 62.9%的試題置屬之，「形成統括和概約理解」和「尋找局部和表層訊息」的試題則分別僅佔 16.13%、3.23%；但「非文學文本題組」則對「發展細部和深層

解釋」、「形成統括和概約理解」、「尋找局部和表層訊息」三個「文本內容」的層次都不偏廢。這顯示「文學文本」和「非文學文本」在閱讀方式上的確有所不同，教學上也應隨不同文本注重不同的閱讀層次。

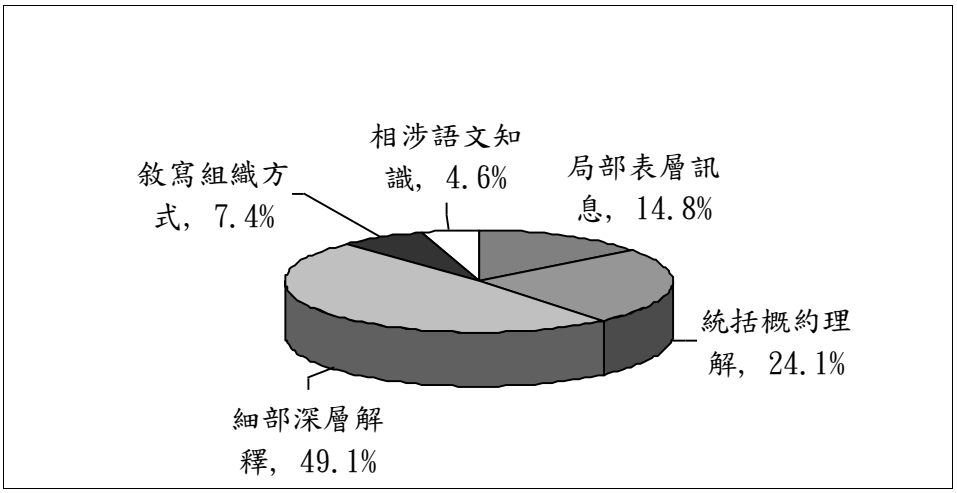


圖5 各閱讀層次的試題數量比

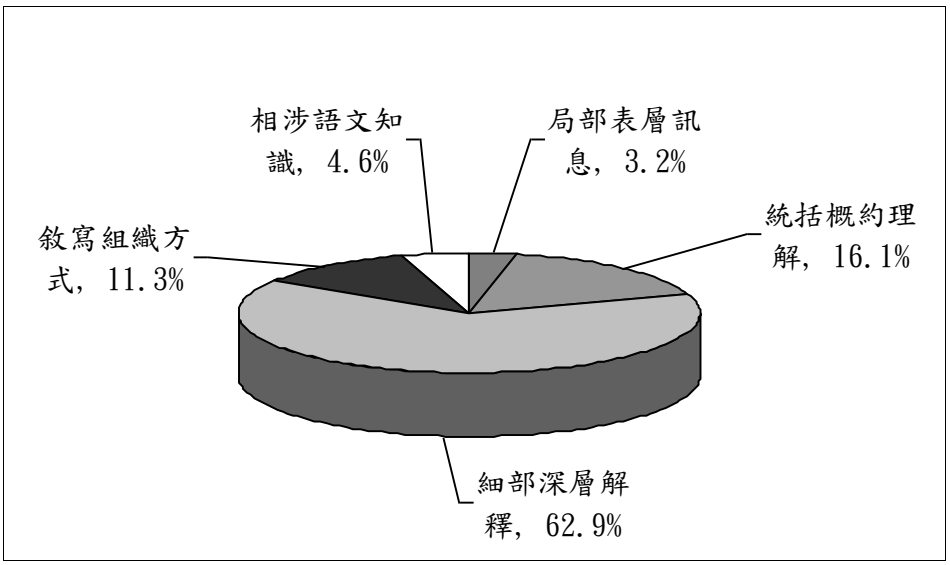


圖6 文學文本試題的閱讀層次分布

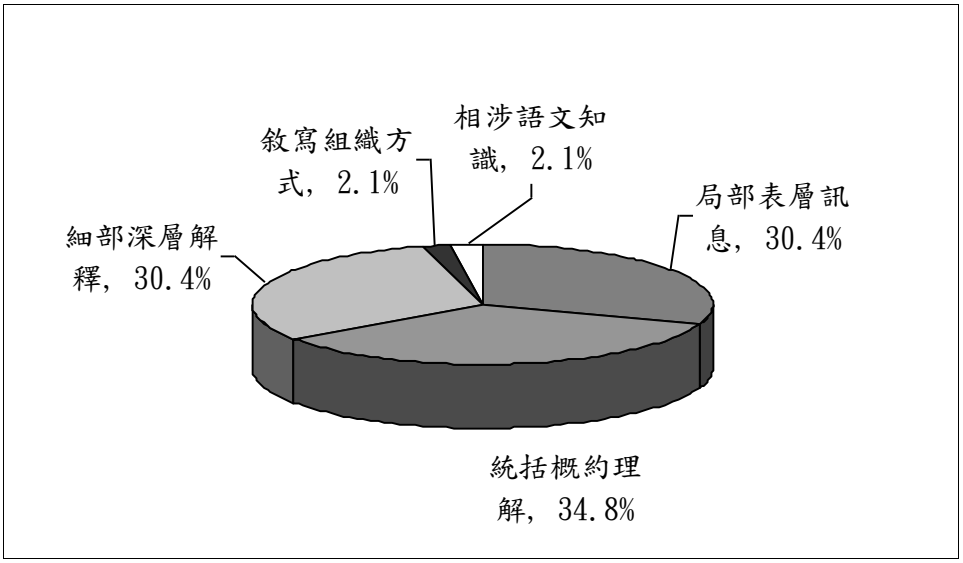


圖7 非文學文本試題的閱讀層次分布

（三）文學文本對考生而言較難；不同閱讀層次的試題難度有別

1、併觀三項考試的國文科「白話文題組」，如圖 8 所示：「文學文本題組」的平均答對率（0.70）低於「非文學文本題組」的平均答對率（0.74），顯然考生較能理解「非文學文本」，較不易理解「文學文本」，而這應與「文學文本題組」明顯偏重「細部和深層的解釋」有關。

2、併觀三項考試的國文科「白話文題組」，分屬五個閱讀層次的試題平均答對率如圖 9 所示，由高至低依序為：「尋找局部和表層訊息」＞「形成統括和概約理解」＞「發展細部和深層解釋」＞「掌握敘寫和組織方式」＞「聯結相涉的語文知識」，足見考生對「文本內容」的理解程度高於對「文本形式」的理解，且在「延伸至文本外」的理解上最容易受挫。

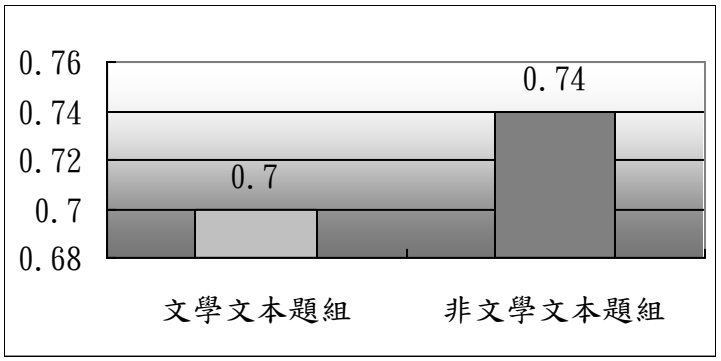


圖8 兩類文本試題的平均答對率

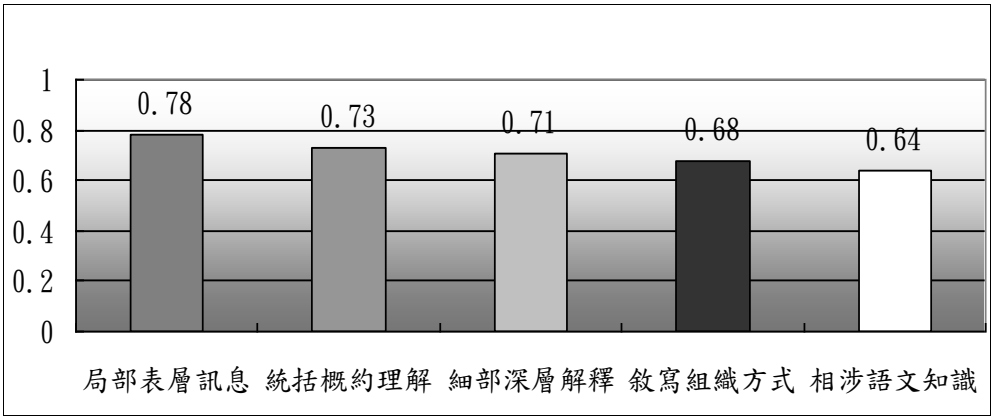


圖9 各閱讀層次試題的平均答對率

二、建議

（一）「非文學文本」也是適合的閱讀能力檢測素材

在現行高中與高職「國文」課程綱要中，固然相當強調「文學作品」是達成下列課程目標的重要橋樑——「增進本國語文聽、說、讀、寫之能力」、「開拓生活視野，關懷生命意義，培養優美情操，提升表達能力」、「與當代環境對話，以理解文明社會之基本價值，尊重多元精神，啟發文化反思能力」（以上高中國文課程綱要）、「培養學生閱讀、表達、欣賞與寫作簡易語體文之興趣及能力」、「陶冶優雅之氣質及高尚之情操」、「培養學生思考、組織、創造及想像之能力」（以上高職國文課程綱要），但「文學」以外的篇章，一則

可與「文學作品」相輔相成，再則實比「文學作品」更為一般人尋常所涉獵，故在學測、指考、統測國文科中以「非文學文本」為閱讀素材來設計試題，不僅未與課程綱要相違，且更能藉由延伸的閱讀觸角、廣泛的閱讀主題，達到課程綱要所設定的諸多「非關文學」的目標。

雖然從考生平均答對率來看，「非文學文本題組」是較「文學文本題組」來得高，但本文的研究也顯示：比較容易答對的「非文學文本題組」，其鑑別度未必遜於「文學文本題組」。故未來編製「白話文閱讀選擇題組」時，實不必對某種素材預存成見，一逕以為「文學文本」絕對優於「非文學文本」。

（二）編製試題無需刻意兼顧各閱讀層次

本文雖然指出：「文本內容」的理解可分為「尋找局部和表層訊息」、「形成統括和概約理解」、「發展細部和深層解釋」三個層次；「掌握敘寫和組織方式」是相對於「文本內容」、屬於「文本形式」的理解；而相對於「集中於文本內」的「內容」和「形式」理解，還有「相涉的語文知識」這層「延伸至文本外」的理解。但在對閱讀素材編寫試題時，並不需要「層層兼顧」。此一則因為「求全」的結果，可能會考到對閱讀素材而言並不重要的層次——例如對「文學文本」來說，「尋找局部和表層訊息」也許可以忽略；對「非文學文本來說」，「文本形式」也許不必細究。另一方面，一份閱讀素材通常也只能有限的配置兩三題，如果拘守「求全」而忘了有時「偏」即可概「全」，反而是浪費。誠如章熊《中國當代寫作與閱讀測試》所云：

> 布魯姆的能力層級學說為教育測量學做出了重要貢獻，……這種思路也延伸到閱讀測試，……閱讀測試既然重視理解的整體性，就勢必重視測試材料的個性，因而每次命題都會帶來一定的靈活變化，不拘泥於原有計畫，不機械地規定各種比例。我國不少者提出，閱讀能力層次中，高層次對於低層次具有一定的覆蓋性，因此比較重視較高層次的測試，通

過較高層次的試題來檢測被試者的閱讀理解能力。（章熊，2000：368）

因此，上述基於文本「內」與「外」、「內容」與「形式」所開展的五個閱讀層次，在編寫試題時，只宜做為「胸有丘壑」的藍圖，不宜完全照章行「試」。

（三）應於選擇題外重視非選擇題

以選擇題進行閱讀能力檢測，原本用來「提問」的試題，其實也會進入「詮釋循環」——亦即當考生理解閱讀素材時，試題的題幹與選項（包含正答與誘答）也會在「自下而上的加工」與「自上而下的加工」中發揮作用，於是便產生「選項把試題變簡單」的情形。例如「緒說」所引 97 年學測國文科考朱熹《朱子語類》的其中一題：

> 上文朱熹以「吃糖」為喻，目的是希望讀書人明白：
> (A)在教學方法上，孔子的身教優於孟子的言教
> (B)孔子說理直截了當；孟子說理繁瑣，言過其實
> (C)孔子雖少講理論，實教人透過生活實踐來體悟道理
> (D)「仁」因孟子的解釋分曉，才確立為儒家學說的核心

朱熹是借「吃糖」來凸顯孔、孟的不同——孟子好言辯，總宣揚「糖是甜的」而要人吃；孔子重實踐，認為只要把糖吃下去就知道糖是甜的。該題的答對率為 90%，然而一旦沒有選項提示，相信答對率不會那麼高。因此，用選擇題考閱讀，便無法避免考生運用「代入、比較、消去」等技巧找答案，考生在閱讀前自認有所憑恃，閱讀時也就不會認真細讀。

或許是鑑於上述閱讀教學的負面因素，近兩三年的學測國文科已在佔全卷題分 50%的「非選擇題」中，撥出 50%編製非選擇式的閱讀題[13]，指考國文科

[13] 99 年學測國文科考王家祥〈秋日的聲音〉、顧炎武〈廉恥〉的解讀；100 年學測國文科考朱光潛〈對於一棵古松的三種態度〉、蘇軾〈赤壁賦〉的解讀；101 年學測國文科金耀基《劍橋語絲》、陶潛〈桃花源記〉的解讀。

也在佔全卷題分 45%的「非選擇題」中，撥出 40%編製非選擇式的閱讀題[14]。相信在大型升學考試命題方向調整的引導下，必能對高中職學生的閱讀能力有所提升。

[14] 98 年指考國文科考〈馮諼客孟嘗君〉的解讀；99 年指考國文科考蔣勳〈關於屈原的最後一天〉的解讀；100 年指考國文科考梁遇春〈途中〉的解讀。

參考文獻

于燕（2006）。NAEP 閱讀評價體系述評。**中學語文教學**，**1**，3-9。

中國教育線上高考頻道（2011）。2011 年全國新課標高考考試大綱：語文。2012 年 2 月 29 日取自 http://gaokao.eol.cn/gkdg_6255/20110307/t20110307_584907.shtml。

天下雜誌教育基金會（2010）。拔尖扶弱——臺灣閱讀教育方向。2012 年 2 月 29 日取自 http://reading.cw.com.tw/doc/page.jspx?id=40288ab22ce7bf65012cf32837a3000b。

余應源（2001）。語文「姓」什麼——認識與從事語文教學的邏輯起點。**中學語文教學**，**3**，9-12。

吳清基（2010）。推動臺灣的閱讀教育——全民來閱讀。**研考雙月刊**，**34**（1），62-66。

范琳、周紅、劉振前（2011）。**二語語篇閱讀推理的心理學研究**。北京：北京大學出版社。

孫明峰、廖純英（2006）。學生閱讀能力評量規準——美國 NAEP 2005 學生閱讀評量測驗之啟示。**大直高中學報**，**3**，75-96。

殷鼎（1990）。**理解的命運**。臺北：東大圖書公司。

張莉慧（2009）。臺灣推動閱讀之觀察與省思。**臺灣圖書館管理季刊**，**5**（4），82-98。

張穎（2006）。國際閱讀素養進展研究（PIRLS）項目評介。**中學語文教學**，**12**，3-9。

章熊（2000）。**中國當代寫作與閱讀測試**。成都：四川教育出版社。

葉連祺、林淑萍（2003）。布魯姆認知領域教育目標分類修訂版之探討。**教育研究月刊**，**105**，94-106。

廖先、祝新華（2010）。從國際閱讀評估項目的最近發展探討閱讀評估策略。**全球教育展望**，**12**，53-59。

臺灣 PISA 國家研究中心（2011）。**臺灣 PISA 2009 結果報告**。臺北：心理出版社。

樂中保（2008）。PISA 中閱讀測試的測評框架與設計思路——兼談對我國閱讀測試的啟示。**河北師范大學學報（教育科學版）**，**10**（6），32-35。

盧雪梅（2011）。國中基測國文科閱讀文本暨學生表現分析。**教育研究與發展期刊**，**7**（2），115-152。

龍協濤（1997）。**讀者反應理論**。臺北：揚智文化事業公司。

學科能力測驗非選擇題閱卷一致性之探討

張銘秋
大學入學考試中心

摘要

學科能力測驗（以下簡稱學測）之國文考科與英文考科皆含需人工閱卷之非選擇題，外界對於非選擇題的閱卷一致性始終非常關心。學測屬於高風險測驗，國內外對此類測驗閱卷一致性的研究文獻不多，因此，有必要進行學測國文考科與英文考科非選擇題閱卷一致性之研究。本研究之目的為探討學測國文考科與英文考科非選擇題的信度，以類推性理論為主，並以評閱分數一致性百分比以及評閱分數間相關輔助說明。期望透過本研究結果，一方面能提供測驗實務界關於高風險考試中如何監控與提升非選擇題閱卷一致性的作法，另一方面可以讓社會相關人士瞭解學測非選擇題閱卷一致性之實徵證據。

關鍵詞：學科能力測驗、閱卷一致性

張銘秋，大學入學考試中心專員

The Rating Consistency of General Scholastic Ability Test

Ming-Chiu Chang
College Entrance Examination Center

Abstract

Rating consistency of The General Scholastic Ability Test (GSAT) has been one of the most concerning issues to public. However, researches on rating consistency of non- multiple-choice items are not sufficient since the GSAT is a large-scale high-stake exam. The purpose of this study is to validate the rating consistency of GSAT. The approaches to validate the rating consistency include generalizability theory and rating consensus. The results show that GSAT, as a large-scale high-stake exam, has high rating consistency. This paper not only presents the evidence on the good rating consistency of GSAT, but it also provides references on how to monitor and improve rating consistency of high-stake non-multiple-choice items.

Keywords: General Scholastic Ability Test, Rating Consistency

---

Ming-Chiu Chang, Staff Member, College Entrance Examination Center

壹、研究動機與目的

學科能力測驗（以下簡稱學測）旨在評量考生是否具有接受大學教育的基本學科能力，是大學校系初步篩選學生的門檻。學測是以電腦可讀的題型為主，例如：選擇題（單選題、多選題）、選填題，而國文與英文兩考科則有需人工閱卷的非選擇題。國文考科的非選擇題，採用「語文表達能力測驗」題型，主要是測驗學生運用文字統整資料、改寫文章、判讀圖表訊息等表達能力；而英文考科的非選擇題則可能包括句子合併、中譯英、英文寫作等（大學入學考試中心，2011）。這些非選擇題的題目設計與計分的特性與一般客觀式試題不同，因此在高風險考試上之實施方式及其結果的信度與效度特別受到矚目。

傳統檢驗信度的方法不外乎三種，即重測、複本、與內部一致性。可是實作評量或檔案評量此類非選擇題式的評量除了有受試者能力因素的考量，尚有閱卷者、作業性質、甚至其他因素的涉入，因此在這個議題上，測驗評量學者或心理計量學家關心的不僅是學生在不同作業間表現的類推性（即能否由受試者的表現推知其能力），不同閱卷者對相同學生表現的閱卷一致性（即不同閱卷者能否對學生表現有相同或類似的解釋與評鑑），甚至是否還有其他來源影響表現分數的一致性，諸如施測時間、受試者所屬單位等因素，於是類推性理論（Generalizability theory，簡稱 G 理論, Brennan, 2001）取而代之成為估計此類評量結果信度的方法（鄒慧英，2004）。

以傳統的閱卷一致性分析法估計信度，在計算上雖然較為簡便，但仍有其限制或缺點。以閱卷者間相關為例，當閱卷者間評閱分數的變異較小，亦即閱卷者間的給分一致性較高時，所得到的相關係數也變小，進而使研究者得到錯誤的結論。再者，這些方法也無從進一步得知閱卷者人數多寡，對測驗分數信度的影響。因此本研究以類推性理論的分析為主，輔以傳統閱卷一致性分析法，探討學測國文考科與英文考科非選擇題的信度。期望透過本研究結果，一方面能提供實務界關於高風險測驗中，如何監控與提升非選擇題測驗信度的作法，

另一方面也可以讓社會人士瞭解學測國文考科與英文考科非選擇題的信度實徵證據。

貳、文獻回顧

一、非選擇題測驗的信度

測驗結果的信度係指測量結果的一致性，其涵義一般可從測量一致性及測量誤差兩個觀點來探討，以測驗一致性的觀點而言，信度係指相同的個人在不同的時間情境下，以相同或複本（相等的試題）測量，所得結果的一致性（consistency）或穩定性（stability）的程度；從測量誤差的觀點，信度是測驗分數中測量誤差所占的比例，誤差所占比例愈低，表示觀察分數愈能反映受試者的真實分數，反之亦然。

在古典測驗理論當中，依此概念所發展的信度估計法包含：重測法、複本法與內部一致性法。其中內部一致性法又可以折半法（split-half）、庫李法（Kuder-Richardson）以及 Cronbach α 法等來進行估計。然而這些方法多適用於無需考量閱卷者變異的客觀式測驗，而學測國文考科與英文考科有需人工閱卷的非選擇題，其涉及閱卷者主觀判斷的特性，因此除上述方法之外，亦需探討閱卷結果之一致性。需人工閱卷之非選擇題閱卷信度估計方法，通常包含閱卷一致性百分比、閱卷者間相關以及類推性理論等三種方法。

閱卷一致性百分比指的是閱卷者評閱相同試卷並給予相同分數的比例。閱卷者相關則是指閱卷者給分間的相關，亦即閱卷者的給分是否有同樣趨高或趨低的模式。在相關係數的計算上，當只有兩位閱卷者時，可以利用 Pearson 積差相關或 Spearman 等級相關；當有三位以上的閱卷者，則可採肯德爾和諧係數（Kendall coefficient of concordance）。Brown、Glasswell 與 Harland（2004）整理了數個大型測驗（如 The New Standards Project, CRESST、Vermont 的檔案評量、英國高風險寫作測驗、以及 ETS 的安置測驗與托福）中寫作測驗的閱卷

一致性。結果發現完全一致的比例通常介於 40%~60%之間，相鄰分數類別的一致性在 80%~100%之間，而相關係數則是介於 0.7~0.8 之間。

上述兩種信度估計法計算簡便且為大部分研究所採用，但在意義的解釋上卻有其限制。以閱卷一致性百分比而言，Novak、Herman 與 Gearhart（1996）認為由於閱卷者所評閱的分數很難達到百分之百的完全一致，因此不同的研究，對於完全一致的比例應達到多少才屬合理，就有不同的標準。在閱卷者間相關方面，由於閱卷者間相關看的僅是閱卷者間評閱分數的相對等級，因此，即使閱卷者間給分的差異較大，但只要給分的高低排序相似，仍然可能得到相當高的相關值。反之，若閱卷者間的給分較為一致，但因為分數間的變異較小、高低排序較為不同，因此將得到較低的相關係數，進而形成錯誤的研究結論（Brown et al., 2004; Novak et al., 1996）。因應閱卷一致性百分比與閱卷者間相關兩種信度估計法在使用上的缺點，學者們開始將類推性理論引入非選擇題試題的信度估計中。

二、類推性理論的基本概念

類推性理論（Generalizability Theory）簡稱 G 理論，提供了確認主要測驗誤差以及估計不同測量程序精確性的概念性和統計的架構（Brennan, 2001）。G 理論可視為古典測驗理論的延伸，因此有一些概念與古典測驗理論是相似的。在古典測驗理論中，一個觀察分數（X）只能分解為一個真分數（true score, T）和一個無法區分的隨機誤差（undifferentiated random error, E），其模式為 X=T+E。但相反的，類推性理論以全域分數（universe score）來替代古典測驗中的真分數，而全域分數指的是受試者在所有測量情境所得到的平均分數之期望值。然而不同於古典測驗理論，類推性理論使用變異數分析（ANOVA）的方法，使研究者不僅能辨別古典測驗理論中的誤差，更能區辨導致誤差的誤差來源（Brennan, 2001）。因此，古典測驗理論與變異數分析就像父母，衍生出類推性理論。然而，類推性理論雖然是古典測驗理論的延伸，但並非所有古典測驗

理論皆能融入此一理論中;同樣的,類推性理論雖然使用了變異數分析的方法，但也不是所有觀點都與變異數分析一致（Brennan, 2001）。Brennan 以圖 1 說明類推性理論與古典測驗理論及變異數分析的關係：

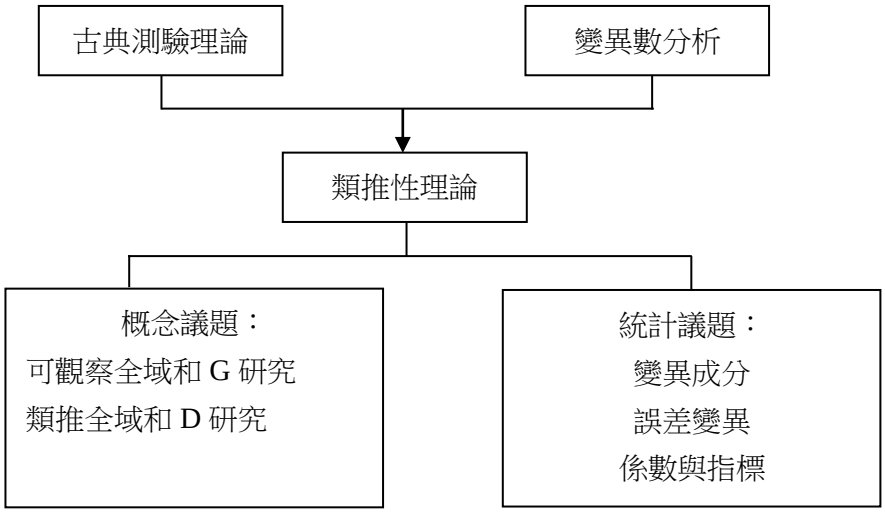


圖 1　類推性理論的源頭與概念架構（Brennan, 2001, p.5）

由圖 1 可以看出類推性理論來自於古典測驗理論與變異數分析，並可從此理論中區分成兩個議題，一為概念議題，另一為統計議題，以下即針對類推性理論的幾個重要議題加以說明：

（一）類推性研究與可觀察全域、決策性研究與類推全域

在了解 G 理論之前，必須先區分類推性研究（Generalizability study, G-study）、可觀察全域（universe of admissible observation）與決策性研究（Decision study, D-study）與類推全域（universe of generalization）。

類推性研究是要提供有關測量誤差來源的訊息，也就是要探討測量樣本（sample of measurements）對測量全域（universe of measurement）可類推的程度。可觀察全域的概念係指類推性研究是在某組特定的情境下進行，這組特定的情境代表著更大群組的情境，在此更大的群組情境中，可觀察的母群便稱為

可觀察全域。舉例來說，研究者在建構測驗時，通常會先確認可能使用的試題，此時研究者不一定會使用任何特定的試題，只是描述感興趣的測量變項，這些測量變項在類推性理論中稱為測量面向（facet）。面向係指一組相似的測量情境（意即研究者關心的測量變項），因此任何一道試題就形成試題面向中一個可接受的測量情境（admissible condition of measurement），此時試題面向就是可觀察全域。由於類推性研究的目的在於協助制訂具有適當類推性的決策性研究，因此，在類推性研究中應該盡可能將可觀全域的範圍訂的廣一點（Brennan, 2001； Shavelson & Webb, 1991），而類推性研究所得的估計值便可用來規劃測量程序，並提供在不同的決策性研究之下，作決策的訊息。

依據類推性研究所提供的訊息，針對某個特定目的設計出的最佳測量方式，以便蒐集資訊作為決策之用，此種研究便稱為「決策性研究」。決策性研究強調的是對變異成分的估計、解釋與使用（Brennan, 2001）。換言之，決策性研究旨在探討測驗設計（如：不同試題數或閱卷者數）的改變，對於欲類推之全域分數信度的影響（Brennan, 2001），也就是依據類推性研究所提供的資訊結果，研究者再依據其目標，調整不同面向內的測量水準數量，以降低誤差或提高類推性係數。

簡言之，類推性研究著重的是對變異成分的界定與估計，而決策性研究強調的則是對變異成分的應用及解釋。兩者除上述之不同外，尚有以下二點差異，以試題數及閱卷者之二面向交叉設計為例說明：1、類推性研究的全域分數（即真分數）是推論回全域中所有單一受試者、單一試題、單一閱卷者上的期望分數或平均分數，其設計的寫法為 $p\times i\times r$，其中 p 代表受試者，即測量對象（objects of measurement）；i 代表試題，r 代表閱卷者，「×」代表交叉；決策性研究的真分數則是類推回全域中單一受試者在 n 個試題、m 位閱卷者的期望分數或平均數（其設計的寫法為 $p\times I\times R$）；2、除受試者之變異成份相同外，其餘則是類推性研究之變異成份的估計均大於或等於決策性研究之變異成份的估計（Brennan, 2001; Shavelson &Webb, 1991）。

（二）固定面向（fixed facet）與隨機面向（random facet）

在 G 理論中面向可視為固定或隨機，當樣本數小於全域的數量，且樣本為隨機抽取或全域中任何相同大小的樣本皆可相互替換時，樣本就被視為隨機（Shavelson & Webb, 1991）。因此，面向是否為隨機仰賴於全域中的情境樣本是否可以替換。舉例來說，若某數學測驗有 25 題，這 25 題皆可由另外的 25 題所替代，則該測驗之試題面向便為隨機面向。反之，當某面向的情境耗盡了類推全域中所有可能的情境，則面向即為固定。舉例來說，數學試題僅能區分成三個認知層次，當數學測驗將這三個認知層次的試題分為三個題組，則題組面向的情境數量就等於類推全域中的情境數量，此時題組這一面向就視為固定面向（Shavelson & Webb, 1991）。

（三）類推性係數（generalizability coefficient）與可靠性指標（index of dependability）

在類推性理論中，有兩種類似信度的係數：類推性係數與可靠性指標，它們代表的都是全域分數變異以及全域分數變異與誤差變異和的比，也就是由受試者的觀察分數類推至全域分數的正確性。但類推係數是由相對決定所得，誤差變異為相對誤差變異，而可靠性指標則是來自絕對決定，誤差變異為絕對誤差的變異。

類推性係數以 $E\hat{\rho}^2$ 來表示，其計算公式為：

$E\hat{\rho}^2 = \dfrac{\hat{\sigma}^2(\tau)}{\hat{\sigma}^2(\tau)+\hat{\sigma}^2(\delta)}$，$\hat{\sigma}^2(\tau)$ 代表全域分數變異，$\hat{\sigma}^2(\delta)$ 代表相對誤差變異。

可靠性指標以 $\hat{\Phi}$ 來表示，其計算公式為：

$\hat{\Phi} = \dfrac{\hat{\sigma}^2(\tau)}{\hat{\sigma}^2(\tau)+\hat{\sigma}^2(\Delta)}$，$\hat{\sigma}^2(\tau)$ 代表全域分數變異，$\hat{\sigma}^2(\Delta)$ 代表絕對誤差變異。

三、大學入學考試中心之閱卷作業

以下分為閱卷者的遴選、閱卷設計以及閱卷委員訓練與閱卷流程，說明大學入學考試中心（以下簡稱大考中心）之閱卷相關作業。

（一）閱卷者的遴選

大考中心閱卷者的聘任模式是採責任制的分層推薦模式。由召集人推薦協同主持人，而在協同主持人會議中推薦閱卷委員，且在確定各科閱卷委員的過程中，閱卷組從閱卷人力庫提供各科歷年閱卷委員之相關資料，及全國各大學相關學系之師資，供各科閱卷召集人及協同主持人參考。各層級的推薦均須經大考中心同意後，方可依各科需求進行意願調查（邱美智、余甄紘，2008）。大考中心對於閱卷編製有詳細的閱卷手冊，其中有關閱卷委員各層級之資格及規約詳見邱美智、余甄紘（2008）。

大考中心依照閱卷者的資格與規約，遴選學測國文考科與英文考科非選擇題的閱卷委員，因此參與的閱卷委員皆為大學教授。每位閱卷委員在參與評分之前，均需經過訓練（內容見閱卷者訓練與閱卷流程一節），以確保閱卷品質。表 1 為 99 與 100 年參與閱卷工作之閱卷委員人數（含正/副召集人與協同主持人）。

表 1　兩年度國文考科與英文考科非選擇題閱卷組數與總閱卷人數

年度		國文考科	英文考科
99	組數	21	12
	人數	231	150
100	組數	18	12
	人數	222	143

（二）閱卷設計

國文考科非選擇題的閱卷部分，大考中心於 1998 年「語文表達能力測驗

研究計畫」設計了以「等級制」取代傳統「百分制」之「三等九級量表」，主要分為 A、B、C 三等，再將各等細分為上、中、下三級，成為三等九級（即 A+、A、A-、B+、B、B-、C+、C、C-）。除「三等九級」的給分方式之外，若出現缺考、未作答、文不對題或作答內容完全照抄試題的情形，則給予 0 分。2010 年由紙面閱卷改採螢幕閱卷模式後，調整成績評定方式，閱卷委員可以在「三等九級」的基礎上，進一步視題旨發揮、資料掌握、結構安排、字句運用等，依題分多寡微調給分。「三等九級量表」的特色在於閱卷委員較能凝聚閱卷共識，使得給分的標準易於掌握而趨於客觀，對閱卷成績評定的品質提升極有助益（大學入學考試中心，2012）。國文考科非選擇題分項式閱卷指標詳見附錄一。

英文考科部分，中譯英每小題各 4 分，原則上是每個錯誤扣 0.5 分。作文的閱卷則是依據內容、組織、文法句構、詞彙拼字與體例各項目分別給分，字數明顯不足則扣總分 1 分。英文非選擇題分項式閱卷指標詳見附錄二。

每位受試者的非選擇題答案由電腦隨機分配予兩位受過訓練的閱卷委員，並由閱卷委員於螢幕上閱卷。當兩位閱卷委員的給分大於事先設定之差分上限時，再由協同主持人進行評閱。表 2 為 99 年與 100 年學測國文與英文考科非選擇題各題題分與需三閱之差分表。

表 2　學測國文與英文考科非選擇題各題題分及需三閱差分表

科目	各題題分及需第三閱差分			
	題號	一	二	三
國文	題分	9	18	27
	差分	>2	>5	>8
英文	題分	8	20	--
	差分	>2	>5	--

（三）閱卷委員訓練與閱卷流程

為使各閱卷委員閱卷標準一致，需先行召開「閱卷標準訂定會議」，由正、

副召集人與協同主持人共同參與，隨機抽取 2000 至 3000 多份受試者答案卷，詳加討論、分析，草擬閱卷標準原則。每題選出各等第之標準卷各 1 份，及試閱卷各 15-20 份。之後再由正、副召集人與協同主持人深入討論、評比所選出的標準卷及試閱卷，並審視、修訂所擬之閱卷原則，確定後，製作閱卷手冊，供正式閱卷前各組協同主持人說明及全體閱卷委員參考之用，並作為閱卷時之參考。

閱卷手冊編製完成後，立即舉辦試閱會議，由召集人統一說明各題閱卷標準及相關注意事項。待閱卷委員詳細閱讀閱卷參考手冊之後進行試閱工作，閱卷委員試閱所有試閱卷後，與各組之協同主持人討論試閱閱卷結果，確認閱卷標準或參考答案，以達成閱卷共識。

之後再由正、副召集人與全體協同主持人舉行閱卷小組會議，目的在於討論試閱會議時閱卷委員提出的問題或新答案，並重新確認閱卷標準或參考答案是否需要更動。之後進行正式閱卷，閱卷時閱卷委員所評閱的前 20 份試卷需經該組協同主持人過目，以確認閱卷委員給分符合閱卷標準；若尚待調整者，由該組協同主持人與其溝通得共識之後，方能進行還卷及接下來的閱卷工作。

正式閱卷前，將由閱卷組與正/副召集人訂定每人總閱卷量與每日閱卷上限，以掌握閱卷品質。一閱委員與二閱委員之分數差距若超過規定之上限，則送請協同主持人進行第三閱，進行第三閱時並不會將前兩閱的分數告知三閱者，以此確保閱卷之合理性與公平性。而從 100 學年度起，增加了第四閱的機制，若第三閱之分數與一、二閱之分數差距大於規定之上限，則由正、副召集人協同另一名資深閱卷委員進行評閱，以減少閱卷的差異。

除了以三閱機制控制閱卷者間的一致性之外，更設置了即時監控系統。在整個閱卷過程中，閱卷委員可隨時查看自己評閱的進度、分數的分布以及與第二閱差分次數統計，並藉由統計結果回饋，以調整其閱卷標準。各組之協同主持人與正、副召集人亦可檢視各閱卷委員之閱卷狀況，若有閱卷委員出現過嚴、過鬆或閱卷不穩定的現象，則可由各組之協同主持人或正、副召集人與閱卷委

員進行溝通，協助閱卷委員釐清問題所在，進而調整其閱卷標準。

為避免閱卷者的疲乏造成閱卷結果不一致，會設定每日閱卷上限份數，其中，上午時段、下午時段與晚上時段各分派定量之試卷。未閱卷完畢的試卷可以延至下個時段再進行閱卷。每位閱卷者可以在限定的工作時間內，自行分配所閱的試卷份數。

參、研究方法

一、資料來源

本研究所使用的資料為 99 與 100 年度學測國文考科與英文考科中的非選擇題試題，並以隨機抽取的 20000 名受試者答題結果進行分析。學測國文考科非選擇題，是採「語文表達能力測驗」題型，主要是考察學生運用文字統整資料、改寫文章、判讀圖表訊息等表達能力。國文考科的測驗題型是在題幹中適度說明，以期藉由說明能更清楚的解釋題意、避免誤解、並幫助受試者多進行聯想、多回憶自己真實感受，如此更能幫助受試者明確鎖定內容，作答更切合題意、避免淪於空洞、表面的論述（大學入學考試中心，2008）。99 與 100 學年度國文考科非選擇題共三大題，第一大題為文章解讀，占分為 9 分；第二大題為文章分析，占分為 18 分；第三大題為引導寫作，占分為 27 分，因此國文考科非選擇題占分共為 54 分。圖 2 為國文考科試題範例。

三、引導寫作

2009年8月，莫拉克颱風所帶來的驚人雨量，在水土保持不良的山區造成嚴重災情，土石流毀壞了橋樑，掩埋了村莊，甚至將山上許多樹木，一路衝到了海邊，成為漂流木。

請想像自己是一株躺在海邊的漂流木，**以「漂流木的獨白」為題，用第一人稱「我」的觀點**寫一篇文章，述說你的遭遇與感想，文長不限。

圖 2　學科能力測驗國文考科非選擇題題目示例

至於學測英文考科非選擇題之目的則是在於評估受試者將中文句子譯成正確、通順與達意之英文能力，以及運用所學詞彙、句法寫出切合主題、並具有統一性與連貫性篇章的能力（大學入學考試中心，2008）。99 與 100 學年度英文考科非選擇題共兩大題，第一大題為文意相連的兩小題中譯英，占分為 8 分；第二大題為英文作文，占分為 20 分，因此英文考科非選擇題占分共為 28 分。圖 3 為英文考科試題範例。

二、英文作文（占 **20** 分）

說明：1.依提示在「答案卷」上寫一篇英文作文。

2.文長至少120個單詞（words）。

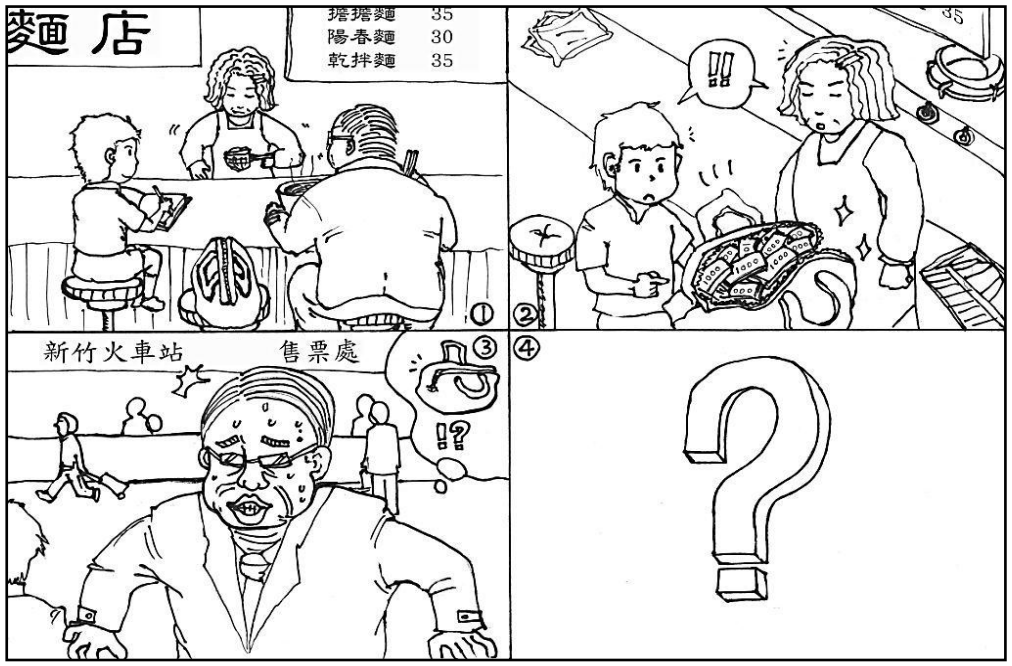


圖 3　學科能力測驗英文考科非選擇題題目示例

二、資料分析

信度的估分法有三種：評閱分數一致性百分比、評閱分數間相關、以及類推性理論。雖然學測非選擇題的閱卷，同一位閱卷委員並非固定都為第一閱或

都為第二閱，在統計分析上不適合以評閱分數間相關來代表評閱分數之一致性。但如前述，本研究中每位閱卷者皆經相同的挑選與訓練，因此，有理由相信每位閱卷者的閱卷標準是相似的，換言之，閱卷者之間可相互替換（interchangeable），意即閱卷者為隨機變項。因此，在此情境之下，仍可以計算兩位閱卷者評閱分數的相關。但由於並非由相同的閱卷者評閱所有受試者的作答，因此本研究以評閱分數（rating）取代閱卷者。

評閱分數一致性百分比的計算，原是直接計算兩個決定同一受試者得分分數中完全一致的比例，以及相差一分以內的比例。但由於學測國文考科與英文考科非選擇題的分數範圍較一般寫作測驗大，再加上原本就設有差分上限，因此本研究將以完全一致、差分在限定範圍之內（不需第三閱），以及需要第三閱的比例，等三種數據來呈現評閱分數一致性百分比。在評閱分數間相關係數部分，則可藉由計算兩個決定受試者得分分數的 Pearson 積差相關來代表。而在類推性理論中，由於類推性理論將個人分數分解成全域分數效果、面向或誤差來源效果，以及每個面向交互作用的效果（Brennan, 2001），因此不同的設計就會產生不同的誤差來源。

在本研究中，由於國文考科的非選擇題為三題，而英文考科為兩題，因此，誤差來源可分為受試者、評閱分數、試題、受試者與評閱分數的交互作用、受試者與試題的交互作用、試題與評閱分數的交互作用，以及三者的交互作用。類推性研究的分析模式可以 $p \times i \times r$ 來表示，此模式原先的意義為每位受試者皆考相同的題目，且由相同的閱卷者進行閱卷，由於受試者為測量對象（母群），並不為面向，而試題與評閱分數為類推全域中可替換的隨機樣本，故採此完全交叉設計模式。在此模式中，μ_p 、μ_i 、 μ_r 等參數均無法直接被估計出，故需以變異數分析的方式估計。受試者 p 在試題 i、第 r 個評閱分數的觀察分數 X_{pir} 之線性模式為：

$X_{pir} = \mu$ 總平均

$+(\mu_p - \mu)$ 受試者效果(正值表示個體分數高過平均數)

$+(\mu_i - \mu)$ 試題效果(正值表示試題難度高過平均數)

$+(\mu_r - \mu)$ 評閱分數效果(正值表示評閱分數高過平均數)

$+(\mu_{pi} - \mu_p - \mu_i + \mu)$ 受試者與試題交互作用效果

$+(\mu_{pr} - \mu_p - \mu_r + \mu)$ 受試者與評閱分數交互作用效果

$+(\mu_{ir} - \mu_i - \mu_r + \mu)$ 試題與評閱分數交互作用效果

$+(X_{pir} - \mu_{pi} - \mu_{pr} - \mu_{ir} + \mu_p + \mu_i + \mu_r - \mu)$.殘差(含受試者,試題與評閱分數交互作用與其他誤差)

每一項效果分配的平均數為 0，變異數為σ^2 。 X_{pir} 分數之總變異量為：

$\sigma^2(X_{pir}) = \sigma_p^2 + \sigma_i^2 + \sigma_r^2 + \sigma_{pi}^2 + \sigma_{pr}^2 + \sigma_{ir}^2 + \sigma_{pir,e}^2$ ，其中$\sigma_{pir,e}^2$為無法再分解之殘差變異。表 3 為 $p \times i \times r$ 二個面向隨機交叉模式的變異數分析摘要表。

表 3 （$p \times i \times r$）二個面向隨機交叉模式變異數分析摘要表

變異來源	離均差平方和	自由度	均方	均方期望值
受試者（p）	SS_p	$n_p - 1$	MS_p	$\sigma_{pir,e}^2 + n_r\sigma_{pi}^2 + n_i\sigma_{pr}^2 + n_i n_r\sigma_p^2$
試題（i）	SS_i	$n_i - 1$	MS_i	$\sigma_{pir,e}^2 + n_r\sigma_{pi}^2 + n_p\sigma_{ir}^2 + n_p n_r\sigma_i^2$
評閱分數（r）	SS_r	$n_r - 1$	MS_r	$\sigma_{pir,e}^2 + n_i\sigma_{pr}^2 + n_p\sigma_{ir}^2 + n_p n_i\sigma_r^2$
$p \times i$	SS_{pi}	$(n_p - 1)(n_i - 1)$	MS_{pi}	$\sigma_{pir,e}^2 + n_r\sigma_{pi}^2$
$p \times r$	SS_{pr}	$(n_p - 1)(n_r - 1)$	MS_{pr}	$\sigma_{pir,e}^2 + n_i\sigma_{pr}^2$
$i \times r$	SS_{ir}	$(n_i - 1)(n_r - 1)$	MS_{ir}	$\sigma_{pir,e}^2 + n_p\sigma_{ir}^2$
$p \times i \times r$	$SS_{pir,e}$	$(n_p - 1)(n_i - 1)(n_r - 1)$	$MS_{pir,e}$	$\sigma_{pir,e}^2$

決策性研究則是根據類推性研究的結果作決策，並推估其信度係數，此決策性研究的模式可以 $p \times I \times R$ 來表示。在決策性研究中，可以透過變異數分解的結果得到類推性係數（ $E\hat{\rho}^2$ 或稱為 G 係數）與可靠性指標（ $\hat{\Phi}$ 或稱為 D 係數），且從類推全域中抽取 n_i' 個試題與 n_r' 個評閱分數，並讓 n_p' 個受試者進行研究，則模式之相對誤差變異量估計與類推性係數，以及絕對誤差變異量與可靠性指標分別如下列方程式所示：

相對誤差變異 $\hat{\sigma}^2(\delta) = \hat{\sigma}_{pI}^2 + \hat{\sigma}_{pR}^2 + \hat{\sigma}_{pIR}^2 = \frac{\hat{\sigma}_{pi}^2}{n_i'} + \frac{\hat{\sigma}_{pr}^2}{n_r'} + \frac{\hat{\sigma}_{pir}^2}{n_i' n_r'}$,

類推性係數 $E\hat{\rho}^2 = \frac{\hat{\sigma}^2(\tau)}{\hat{\sigma}^2(\tau) + \hat{\sigma}^2(\delta)} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_{pi}^2}{n_i'} + \frac{\hat{\sigma}_{pr}^2}{n_r'} + \frac{\hat{\sigma}_{pir}^2}{n_i' n_r'}}$,

絕對誤差變異 $\hat{\sigma}^2(\Delta)$

$= \hat{\sigma}_I^2 + \hat{\sigma}_R^2 + \hat{\sigma}_{pI}^2 + \hat{\sigma}_{pR}^2 + \hat{\sigma}_{IR}^2 + \hat{\sigma}_{pIR}^2 = \frac{\hat{\sigma}_i^2}{n_i'} + \frac{\hat{\sigma}_r^2}{n_r'} + \frac{\hat{\sigma}_{pi}^2}{n_i'} + \frac{\hat{\sigma}_{pr}^2}{n_r'} + \frac{\hat{\sigma}_{ir}^2}{n_i' n_r'} + \frac{\hat{\sigma}_{pir}^2}{n_i' n_r'}$,

可靠性指標 $\hat{\Phi} = \frac{\hat{\sigma}^2(\tau)}{\hat{\sigma}^2(\tau) + \hat{\sigma}^2(\Delta)} = \frac{\hat{\sigma}_p^2}{\hat{\sigma}_p^2 + \frac{\hat{\sigma}_i^2}{n_i'} + \frac{\hat{\sigma}_r^2}{n_r'} + \frac{\hat{\sigma}_{pi}^2}{n_i'} + \frac{\hat{\sigma}_{pr}^2}{n_r'} + \frac{\hat{\sigma}_{ir}^2}{n_i' n_r'} + \frac{\hat{\sigma}_{pir}^2}{n_i' n_r'}}$.

上述信度估計法中所使用的兩個分數來自決定受試者得分的閱卷者給分。在本研究中，當兩個閱卷者的給分在差分上限之內，則採用兩個閱卷者所評閱的分數；當兩個閱卷者給分大於差分上限時，則送交協同主持人進行第三閱，若第三閱評定的分數介於一、二閱之間，則所取的兩個分數皆為第三閱所評閱的分數；若第三閱評定的分數不介於一、二閱之間，則所選取的兩個分數為第三閱的給分，以及與其較為接近的分數。而從 2010 年開始，若第三閱之分數同時與一、二閱之分數差距大於規定之上限，則需進行第四閱，由正、副召集人與一名資深閱卷者共同決定受試者之得分，此時選取的兩個分數皆為第四閱

所評定之分數。以國文考科第三題為例，當兩個閱卷委員給分分別為 10 與 12 分，估計信度時所取的兩個分數即為 10 與 12；而當兩個閱卷者給分分別為 4 與 14，兩者差分大於差分上限，而第三閱給分為 16 分時，所取的兩個分數為 14 與 16 分，若第三閱的給分為 10 分，則所取的兩個分數皆為 10 分，若第三閱給分為 2 分，則所取的兩個分數為 4 分與 2 分，若第三閱給分為 25 分，則需送交第四閱，此時所取的兩個分數皆為第四閱的分數。決定受試者兩個評閱分數之後，以 GENOVA（Crick & Brennan, 1983）進行類推性研究與可靠性研究的分析。

肆、研究結果

學科能力測驗之國文考科與英文考科非選擇題的信度可分成評閱分數一致性百分比、評閱分數間相關以及類推性理論等三部分，以下分別就各部分加以說明。

一、評閱分數一致性百分比

如前述，由於學測國文考科與英文考科非選擇題的分數範圍較一般寫作測驗大，再加上原本就設有差分上限，因此在本研究將以完全一致、差分在限定範圍之內（不須第三閱），以及需要第三閱的比例，等三種數據來呈現評閱分數一致性百分比。表 4 為 99 與 100 年度學測國文考科與英文考科非選擇題各題之評閱分數一致性百分比。

由表中可知除英文考科的中譯英之外，其餘試題評閱分數完全一致之百分比都低於 Brown 等人（2004）指出的 40%至 60%。但在 Brown 等人所整理的大型寫作測驗多為六點或七點計分，而學測的非選擇題最少為中譯英的 9 點計分（0 分亦為一個分數類別），最多則為引導寫作的 28 點計分。學測非選擇題的可能分數類別遠遠大於 Brown 等人研究中所整理的大型寫作測驗，因此作者認為在試題可能分數類別數較多的情境之下，兩個評閱分數完全一致的比例較

低相當合理。表 4 中英文考科的完全一致數據亦可以印證，例如 99 年中的中譯英為 50.3%；作文為 25.4%，前者的一致性大約為後者的二倍，乃因前者占分為 8 分，後者則為 20 分所致；同樣的情形也出現在 100 年的英文考科中。

再者，若加上一、二閱差異在差分範圍內的比例，則除 99 年國文考科的文章解讀之外，兩個評閱分數的一致性都在 90%以上，顯示評閱分數間具有相當高的一致性。

表 4　學測國文考科與英文考科非選擇題評閱分數間一致性百分比

年度	一致性類別	國文考科			英文考科	
		文章解讀	文章分析	引導寫作	中譯英	作文
99	完全一致	26.6%	32.9%	33.6%	50.3%	25.4%
	差分內	61.7%	58.7%	61.0%	48.4%	71.0%
	第三閱	11.7%	8.4%	5.4%	1.3%	3.5%
100	完全一致	27.8%	19.3%	11.7%	46.7%	23.3%
	差分內	63.4%	74.1%	85.0%	51.3%	72.8%
	第三閱	8.8%	6.6%	3.4%	2.0%	4.0%

二、評閱分數間相關

評閱分數間相關係數部分，則計算兩個決定受試者得分分數的 Pearson 積差相關（r），其結果呈現於表 5。由表中可以看出除 100 年國文考科文章解讀（r=0.689）之外，兩年度非選擇題評閱分數的積差相關係數皆在 0.7 以上，符合 Brown 等人（2004）指出相關係數介於 0.7 至 0.8 的標準，其中英文考科的中譯英兩個評閱分數的相關更高達 0.9 以上，顯示學測國文與英文考科非選擇題兩個評閱分數之間的評閱結果非常一致。

表 5　學測國文考科與英文考科非選擇題評閱分數間之相關係數

相關係數	年度	國文考科			英文考科	
		文章解讀	文章分析	引導寫作	中譯英	英文作文
Pearson's *r*	99	.746	.846	.778	.901	.847
	100	.689	.809	.706	.916	.881

三、類推性研究

由於國文考科與英文考科非選擇題的配分不一，因此若直接以原始得分進行分析，將會因為分數量尺不同導致作業之變異大幅增加。因此，在本研究中將原始得分轉換為該題之得分百分比，亦即將原始得分除以該題滿分，並以轉換後分數進行類推性研究之分析。表 6 為兩年度國文與英文考科非選擇題受試者原始得分與得分率之平均數與標準差，由表中可知兩年度國文考科閱卷委員在各題的平均評閱分數相當接近，而在英文考科部分則是 100 年之平均評閱分數略高於 99 年。

學測國文考科與英文考科非選擇題類推性研究之結果呈現於表 7 與表 8。在類推性理論 $p \times i \times r$ 的模式中，受試者變異如同於古典測驗理論中的真分數概念，而閱卷者、試題以及三者之間所有交互作用的變異為誤差。因此，當受試者變異成分所占比例越高，而其他變異成分比例越低時，代表受試者的得分越能反映其真實能力。

表 6　兩年度學測國文與英文考科非選擇題之描述統計

考科	題號	分數類別	描述統計	99 年	100 年
國文	文章解讀	原始分數	平均數	5.18	5.23
			標準差	1.48	1.48
		得分率	平均數	.58	.58
			標準差	.16	.16
	文章分析	原始分數	平均數	7.66	7.31
			標準差	3.27	3.67
		得分率	平均數	.43	.41
			標準差	.18	.20
	引導寫作	原始分數	平均數	11.66	12.28
			標準差	4.37	4.12
		得分率	平均數	.43	.45
			標準差	.16	.15
英文	中譯音	原始分數	平均數	3.09	4.16
			標準差	2.10	2.37
		得分率	平均數	.39	.52
			標準差	.26	.30
	英文作文	原始分數	平均數	6.89	8.03
			標準差	4.55	4.66
		得分率	平均數	.34	.40
			標準差	.23	.23

表 7 為兩年度國文考科非選擇題 $p \times i \times r$ 類推性研究的結果，從表中可知兩年度資料中變異最大的部分皆為受試者（p），其次皆為受試者與試題之交互作用（$p \times i$）。兩年度類推性研究結果皆顯示評閱分數（r）的變異為 0，顯示 2 個評閱分數間的閱卷結果非常一致，而兩年度試題與評閱分數間交互作用（$i \times r$）亦為 0，顯示 2 個評閱分數間不易受試題之影響。兩年度第二大之變異來源為受試者與試題之交互作用（$p \times i$），占總變異之 28%以上，顯示受試

者在不同試題有不同的表現。

表 8 為英文考科的類推性研究結果，從表中可知最大變異來源為受試者，兩年度受試者之變異皆占總變異之 70%以上，其餘變異皆不大。與國文考科相同，2 個評閱分數間變異皆為 0，顯示閱卷者間之閱卷結果非常一致，而試題與評閱分數間交互作用（$i \times r$）之變異亦為 0，顯示閱卷者評閱分數不易受到試題之影響。

表 7　學測國文考科非選擇題 $p \times i \times r$ 類推性研究結果

變異來源	99 年			100 年		
	變異量	自由度	百分比	變異量	自由度	百分比
p	.0158	19999	39.74%	.0140	19999	32.09%
i	.0051	2	12.85%	.0082	2	18.70%
r	.0000	1	0%	.0000	1	0%
$p \times i$	.0112	39998	28.07%	.0129	39998	29.42%
$p \times r$	.0015	19999	3.78%	.0018	19999	4.08%
$i \times r$	.0000	2	0%	.0000	2	0%
$p \times i \times r$	.0062	39998	15.56%	.0069	39998	15.71%

表 8　學測英文考科非選擇題 $p \times i \times r$ 類推性研究結果

變異來源	99 年			100 年		
	變異量	自由度	百分比	變異量	自由度	百分比
p	.0492	19999	75.23%	.0594	19999	72.79%
i	.0007	1	1.08%	.0070	1	8.58%
r	.0000	1	0%	.0000	1	0%
$p \times i$	.0070	19999	10.70%	.0089	19999	10.91%
$p \times r$	.0022	19999	3.36%	.0012	19999	1.47%
$i \times r$	.0000	1	0%	.0000	1	0%
$p \times i \times r$	.0063	19999	9.63%	.0051	19999	6.25%

圖 4 與圖 5 為兩年度國文考科非選擇題之決策性研究分析結果。由圖中可以看出評閱分數個數、試題數與類信度係數值間的關係。國文考科類推性研究

結果雖然顯示受試者之變異最大，但受試者與試題之交互作用之變異也很大，因此在目前2個評閱分數與3題試題之下，99年國文考科之類推性係數為0.71、100 年為 0.69。由於與評閱分數相關的變異相對較小，因此增加評閱分數個數所能提升的類推性係數較為有限（.023-.050），而增加試題數的效果與增加評閱分數個數的效果相似。根據決策性研究的結果顯示 99 年需增加為 3 個評閱分數與 5 個試題，而 100 年則需增加為 3 個評閱分數與 6 個試題，才能達到 Shavelson 與 Webb（1991）所設定的 0.80 以上的類推性係數。

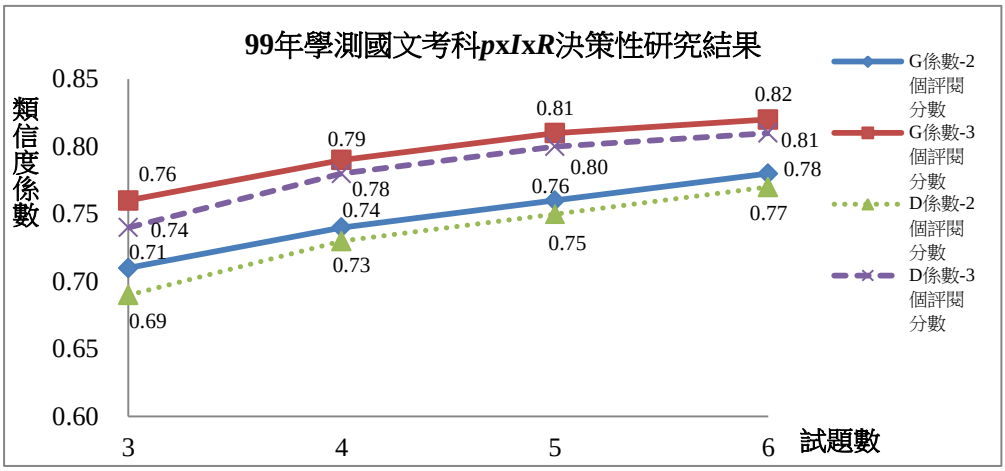


圖 4　99 年國文考科 $p \times I \times R$ 決策性研究結果圖示

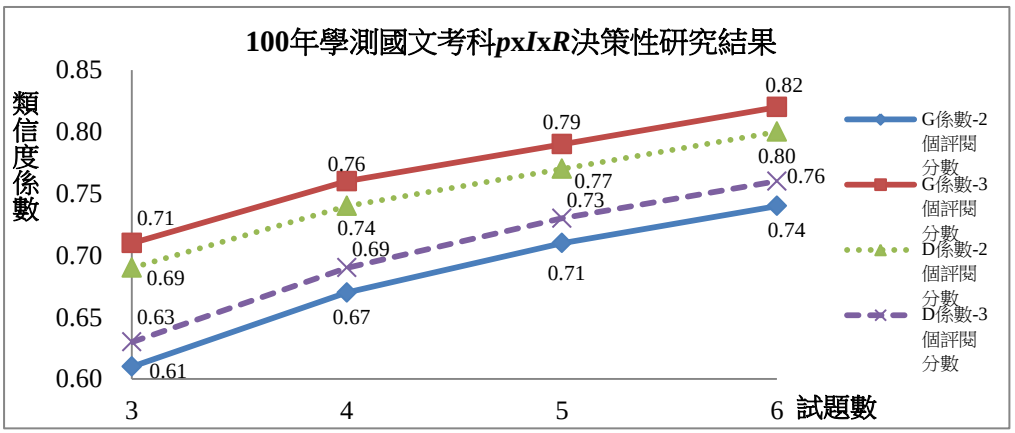


圖 5　100 年學測國文考科 $p \times I \times R$ 決策性研究結果圖示

圖 6 與圖 7 則為兩年度英文考科非選擇題之決策性研究分析結果。由這兩個圖可以看出，當閱卷者人數為現行兩人且試題數亦為現行之兩題時，99 年的類推性係數已達 0.89，100 年則為 0.90，顯示兩年度英文考科非選擇題之閱卷具有良好的信度。再者，結果也顯示學測英文科僅需 2 個評閱分數與 1 個試題，即可達到 Shavelson 與 Webb（1991）所設定的.80 以上的類推性係數，再次顯示英文考科非選擇題之閱卷相當一致。

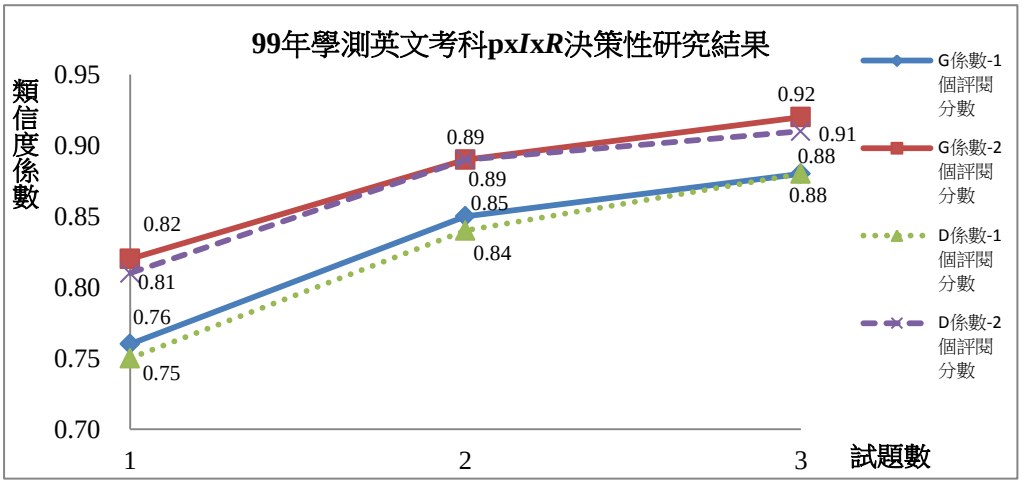


圖 6　99 年學測英文考科 $p \times I \times R$ 決策性研究結果圖示

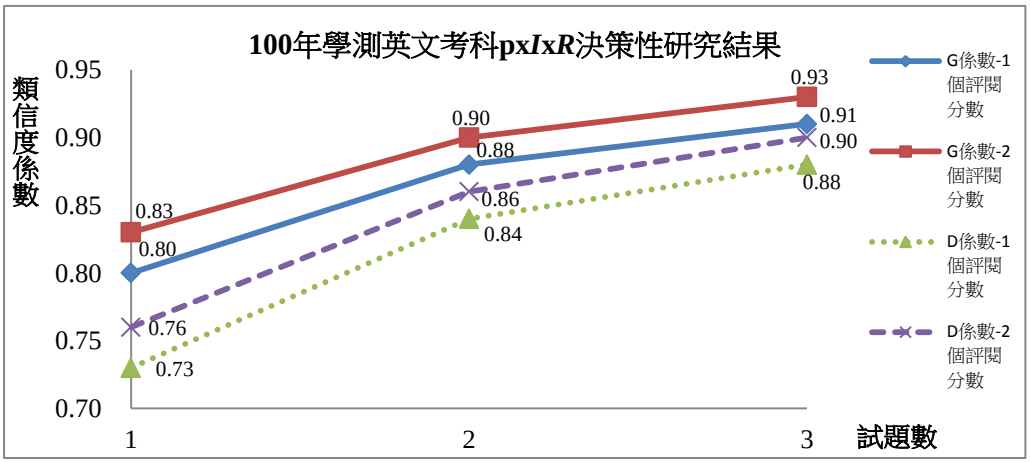


圖 7　100 年學測英文考科 $p \times I \times R$ 決策性研究結果圖示

作者亦發現國文考科非選擇題試題變異頗大，兩年度試題之變異占總變異10%以上，受試者與試題間交互作用($p \times i$)之變異更是占了近 30%的總變異；而英文考科非選擇題的試題變異則相對較小，但 $p \times i$ 的變異也占了 10%以上。因此，作者進一步檢視兩年度國文考科與英文考科非選擇題各題得分間的相關係數，結果如表 9 所示。對照於英文考科兩題非選擇之相關係數，99 年為.839，100 年為.866，可知國文考科三題非選擇題之間的相關係數較低，最高為 100 年第二題與第三題的相關，但也僅為.575。因此，作者推測原因可能在於設計試題時，國文考科三題非選擇題所欲測量的能力有所不同之故。

學測國文考科非選擇題是採「語文表達能力測驗」題型，也就是希望受試者能藉由一段敘述的觸發，進行選取素材、組合成文的工作。然而，寫作不僅可能會針對不同的場合選擇不同的文體，完整的寫作任務也是由好幾個階段組合而成。因此，在語文能力的考查上，除了有不同寫作題型如抒情、議論、人物刻畫或景物描寫等等之外，也會有不同的寫作要求如擴寫、接寫、修飾與摘要等等。且國文考科三題非選擇題的測驗功能皆不同，第一題著重於知性的統整判斷能力，要求受試者歸納與分析資料；第二題著重於基本語文應用能力，受試者須以準確清晰的語言陳述自己的立場；第三題則是評量情意的感受抒發能力（大學入學考試中心，2008）。再者，多數受試者皆有本身擅長與不擅長的寫作題型與要求，因此在不同題型與要求上的表現可能會不一致，因而導致受試者與試題之交互作用有較大的變異。英文考科的二題非選擇題中「中譯英」主要在於測量受試者能否將中文句子譯成正確、通順、達意之英文；而「作文」則是評量受試者能否依據提示，運用所學詞彙、句法寫出切合主題，並具連慣性之短文（大學入學考試中心，2008）。「中譯英」的正確、通順、達意之英文，較類似於「作文」分項式閱卷指標中的「文法、句構」與「體例」兩項目。因此，雖然英文考科非選擇題測量的目標略有不同，但相較於國文考科的非選擇題而言，其相似性仍較高。

表 9　兩年度國文與英文考科非選擇題間相關係數

考科	年度	題號	第一題	第二題
國文	99	第二題	.540	--
		第三題	.431	.519
	100	第二題	.478	--
		第三題	.418	.575
英文	99	第二題	.839	--
	100	第二題	.866	--

伍、結論

本研究之目的在於瞭解學科能力測驗國文考科與英文考科中非選擇題的閱卷一致性，以下分別歸納研究結果。

本研究係以不同統計方法進行分析。由分析結果發現，評閱分數之一致性已達到令人滿意的程度。顯示事前的閱卷者訓練、評選與嚴謹的評閱流程，能使學測國文考科與英文考科的非選擇題具有良好的閱卷結果。

在類推性理論當中，英文考科以 *p*×*i*×*r* 模式進行分析即可得到不錯的類信度係數，且與評閱分數有關的變異都很小或接近於 0，顯示學測英文考科非選擇題的閱卷非常一致，且僅需 2 個評閱分數與 1 題試題，便可得到 0.80 以上之類推性係數。而同樣以 *p*×*i*×*r* 模式分析國文考科非選擇題時，雖然與評閱分數有關的變異也都很小，但由於受試者與試題交互作用($p \times i$)之變異相對較大，因此其類信度係數較英文考科為低，且需 3 個評閱分數與 6 題以上的試題才能達到 0.80 以上之類推性係數。

需注意的是兩考科兩年度第二大之變異來源皆為受試者與試題之交互作用($p \times i$)，國文考科之變異占總變異之 28%以上，而英文考科則占 10%以上。作者推測原因在於非選擇題的功能不一，且多數受試者皆有本身擅長與不擅長的寫作題型與要求，在不同題型與要求上的表現可能會不一致，進而導致受試者與試題之交互作用有較大的變異。此外，由於閱卷時並沒有限制必須採取分

題閱卷方式，因此，多數閱卷委員皆以受試者為單位，評閱完一位受試者之所有試題之後，再評閱下一位受試者之答題反應。如此作法，容易使閱卷者評閱同一位受試者之其他試題時，受到前一個試題的影響。或是因為評閱不同試題需使用不同評分標準，所以未採取分題閱卷方式，除不易維持閱卷標準之一致性之外，對於閱卷者之認知負荷也較大。因此，建議可改採分題閱卷之方式，唯若採分題閱卷方式，則包含閱卷設計、閱卷流程、電腦閱卷程式，甚至是答案紙設計與掃瞄等等皆須重新設計，牽涉層面甚廣，因此需進一步探討採用分題閱卷方式之成本與效益的分析。

由閱卷者訓練與閱卷流程一節，可知大考中心對閱卷者的訓練相當嚴謹，但除在正式閱卷訓練之外，閱卷者並沒有其他閱卷的訓練。因此，建議可參考國民中學學生基本能力測驗（簡稱國中基測）中寫作測驗的作法。在國中基測的寫作測驗中，除了例行之樣卷會議（類似於閱卷標準訂定會議）之外，核心評分者（類似於協同主持人）每年尚須參加認證測試及大型樣卷會議。認證測試之目的為確認核心評分者評分標準的穩定性，認證時，核心評分者被隨機分配予該學期討論過的題目 2 題，每題批改 150 份作品。批改結束後，以試題反應理論（item response theory）的 Rasch 模式檢驗評分者是否有過嚴、過鬆或評分不穩定的現象。每位核心評分者皆須通過檢驗，始具有於正式測驗閱卷的工作資格，如此便可確認評分的穩定性，亦為篩選核心評分者之指標。對於一般的評分者（類似閱卷委員），則是一年進行兩次的訓練，使每位參與評分的評分者都能夠充分理解評分規範並修正評分標準（王德蕙、李奕璇、曾芬蘭、宋曜廷，2010）。

參考文獻

大學入學考試中心（2008）。**認識學科能力測驗 9**。臺北：大學入學考試中心。

大學入學考試中心（2011）。**學科能力測驗簡介**。臺北：大學入學考試中心。

大學入學考試中心（2012）。**學測與指考英文、國文作文分項式閱卷指標簡介**。臺北：大學入學考試中心。

王德蕙、李奕璇、曾芬蘭、宋曜廷（2010）。**國民中學學生基本學力測驗寫作測驗信度與效度分析研究**。發表於第九屆海峽兩岸心理與教育測驗暨 2010 NAER 永續教育發展-創新與實踐國際學術研討會，臺北市。

邱美智、余甄紘（2008）。大學入學考試中心紙面閱卷的發展及作業流程。**考試學刊**，**4**，161-186。

鄒慧英（2004）。讀寫檔案的信度與評分者一致性。載於臺南師範學院主編：**「科技化測驗與能力指標評量」國際學術研討論文集**。臺南：臺南師範學院，337－368頁。

Brennan, R. L. (2001). *Generalizability theory*. New York: Springer.

Brown, G. T. L., Glasswell, K., & Harland, D. (2004). Accuracy in the scoring of writing: Studies of reliability and validity using a New Zealand writing assessment system. *Assessing Writing*, *9*, 105-121.

Crick, J. E., & Brennan, R. L. (1983). *Manual for GENOVA: A generalized analysis of variance system* (American College Testing Technical Bulletin No. 43). Iowa City, IA: ACT, Inc.

Novak, J. R., Herman, J. L., & Gearhart, M. (1996). Establishing validity for performance-based assessments: An illustration for collections of student writing. *The Journal of Educational Research*, *89*(4), 220-233.

Shavelson, R. J., & Webb, N. M. (1991). *Generalizability theory: A primer*. Newbury Park: SAGE.

附錄一 國文考科作文分項式閱卷指標

等級 項目	A 等 （A+、A、A-）	B 等 （B+、B、B-）	C 等 （C+、C、C-）
題旨發揮 （40%）	1. 能掌握題幹求，緊扣題旨發揮 2. 內容充實，思路清晰 3. 感發得宜，想像豐富 4. 情感真摯，表達適當，體悟深刻 5. 論述周延，富有創意	1.尚能掌握題幹要求，依照題旨發揮 2.內容平實，思路尚稱清晰 3.感發尚稱得宜，想像平淡 4.情感表達尚稱適當，體悟稍欠深刻 5.論述尚稱周延，略有創意	1. 未能掌握題幹要求，題旨不明或偏離題旨 2. 內容浮泛，思路不清 3. 感發未能得宜，想像不足 4. 情感表達不當，體悟膚淺或全無體悟 5. 論述不周延，缺乏創意
資料掌握 （20%）	1. 能融會貫通題幹資料 2. 能深刻回應引導內容 3. 能善用成語及典故 4. 舉證詳實貼切 5. 材料運用恰當	1. 僅側重部分題幹資料 2. 僅大致回應引導內容 3. 尚能運用成語及典故 4. 舉證平淡疏略 5. 材料運用尚稱恰當	1. 誤解題幹資料 2. 大部分抄襲引導內容 3. 錯誤運用成語及典故 4. 舉證鬆散模糊 5. 材料運用不當
結構安排 （20%）	1. 結構嚴謹 2. 前後通貫 3. 脈絡清楚 4. 條理分明 5. 照應緊密	1. 結構大致完整 2. 前後尚能通貫 3. 脈絡大致清楚 4. 條理尚稱分明 5. 略有照應	1. 結構鬆散 2. 前後矛盾 3. 脈絡不清 4. 條理紛雜 5. 全無照應
字句運用 （20%）	1. 字句妥切，邏輯清晰 2. 用詞精確，造句工穩 3. 描寫細膩，論述精彩 4. 文筆流暢，修辭優美 5. 標點符號使用正確	1.字句尚稱適當，邏輯尚稱清晰 2. 用詞通順，造句平淡 3. 描寫平淡，論述平實 4. 文筆平順，修辭尚可 5. 標點符號使用大致正確	1. 字句欠當，邏輯不 2. 用詞粗率，造句冗贅 3. 描寫粗陋，論述空洞 4. 文筆蕪蔓，修辭粗俗 5. 標點符號使用多有錯誤

資料來源：大學入學考試中心（2012）

附錄二 英文考科作文分項式閱卷指標

等級 項目	優	可	差	劣
內容	主題（句）清楚切題，並有具體、完整的相關係細節支持。 （5-4 分）	主題不夠清楚或凸顯，部分相關敘述發展不全。 （3 分）	主題不明，大部分相關敘述發展不全或與主題無關。 （2-1 分）	文不對題或沒寫（凡文不對題或沒寫者，其他各項均以零分計算）。 （0 分）
組織	重點分明，有開頭、發展、結尾，前後連貫，轉承語使用得當。 （5-4 分）	重點安排不妥，前後發展比例與轉承語使用欠妥。 （3 分）	重點不明，前後不連貫。 （2-1 分）	全文毫無組織或未按提示寫作。 （0 分）
文法、句構	用字精確、得宜，且幾無拼字錯誤。 （5-4 分）	用字單調、重複，用字偶有不當，少許拼字錯誤，但不影響文意之表達。 （3 分）	用字、拼字錯誤多，明顯影響文意之表達。 （2-1 分）	只寫出或抄襲與提議有關的零碎字詞。 （0 分）
體例	格式、標點、大小寫幾無錯誤 （2 分）		格式、標點、大小寫等有錯誤，但不影響文意之表達。 （1 分）	違背基本的寫作體例或格式，標點、大小寫錯誤甚多。 （0 分）

資料來源：大學入學考試中心（2012）

高中階段之地球科學迷思概念與試題探討

翁群評
大學入學考試中心

摘要

地球科學是一門生活化的學科，各種自然現象不斷在生活中上演。因為息息相關，加上人類與生俱來的好奇心與求知慾，多數人對這些自然現象的初始概念，可能是來自於口耳相傳或自我探索。但在這樣的過程中，卻容易形成錯誤認知。

在進入求學階段後，學生並不是腦袋空空的進入學習情境，他們在學習之前，已經存有許多自己的想法，這些先入為主的概念，通常有別於專家的科學概念，且非常難以改變，我們稱之為迷思概念（misconception）。雖然透過正規教育可以慢慢導正這些迷思概念，但其歷程並非平順，學生可能需要不斷重複的學習，才能取代以往的錯誤認知。

教學不止是將新資訊帶入學生現有的知識中而已，而是要讓學生能夠將新知識與舊有的知識架構，統整後再重新加以建構。而測驗是教學的一部分，不是獨立在教學之外的，不同的求學階段，試題測驗仍扮演著重要角色。若能在試題設計上多下點功夫，無論是採題幹上的直接敘述（開門見山）或選項設計上的推理思考(迂迴轉進)，不斷在試題中融入這些學生容易產生的迷思概念，除了達到提醒學生的目的外，可以慢慢產生迷思概念上的釐清，是此研究的探討目的。

關鍵詞：地球科學、迷思概念、試題、測驗、高中

翁群評，大學入學考試中心專員

Misconceptions and Test-Questions of Earth Science in Senior High School

Chun-Ping Weng
College Entrance Examination Center

Abstract

Earth Science is a subject highly related to everyday life because people have to face a variety of natural phenomena constantly in their lives. People might want to know more about Earth Science because of the close relation between natural phenomena and life and the born curiosity and thirst for knowledge in mankind. However, it is likely to develop a wrong perception since many of people’ s initial concepts of these phenomena come from word of mouth or self-exploration.

Students do not come to school with their heads empty. They have many of their own ideas before they acquire knowledge. These preconceptions, often misconceptions and very difficult to change, are usually different from the concepts of scientific experts. Though these misconceptions can be adjusted through formal education, students may need to experience a long process of repeated learning in order to replace those wrong concepts with the correct ones.

Education not only adds new information to students' existing knowledge, but also helps students reestablish their knowledge by combining new and old information. Testing is a part of education and should play an important role in different stages of learning. The purpose of this research is to design functional test items, which are integrated directly or indirectly with common misconceptions, to help students clarify these misconceptions while they are trying to solve the test questions related to Earth Science.

Keywords: Earth Science, Misconceptions, Test Questions, Senior High School

---

Chun-Ping Weng, Staff Member, College Entrance Examination Center

壹、前言

地球科學是一門生活化的學科，因為各種自然現象不斷的在日常生活中上演，例如日月星辰的東昇西落、天氣的變化多端、地形地貌的變遷和海水的潮起潮落等，這都是生活中的一環。因為息息相關，加上人類與生俱來的好奇心與求知慾，多數人對這些自然現象的初始概念，可能是來自於口耳相傳或自我探索。但在這樣的過程中，卻容易形成錯誤認知。

許多探討學生概念理解的研究，指出學生並不是腦袋空空的進入學習情境。他們在學習之前，就已經存有許多自己的想法，這些先入為主的概念，對於他們的學習，有相當大的干擾。當新學習的概念，與他們先前存有的概念相衝突時，自然而然會產生一股抗衡的力量，而兩相衝突之下，當然會對學習的成果有所不利。這個學生先入為主的概念，通常有別於專家的科學概念，且非常難以改變。因此，陳淑筠（2002）指出凡是指學生從生活經驗、學校學習、同儕文化…等中所得，用以了解和解釋自然現象的一套想法，亦即在接受教學前自行建構的與科學家所持有的科學概念相左的概念。凡有別於一般公認之「專家概念」者，我們稱之為迷思概念（misconception）。

雖然透過正規教育可以慢慢導正這些迷思概念，但其歷程並非平順，學生可能需要不斷重複的學習,才能取代以往的錯誤認知。對於如何教導迷思概念，國內外已有相當的多的相關研究。本研究想要探討的是，如何透過測驗中試題設計，引起學生對這些迷思概念的重視與瞭解，進而能在學習過程中，盡早破除這些迷思概念。

測驗是教學的一部分，不是獨立在教學之外的。適當的作業（也是一種評量）與測驗，能夠讓教師和學生認識與了解「教」與「學」訊息，以便調整「教」與「學」的方式。郭重吉（1988）指出教學不止是將新資訊帶入學生現有的知識中而已，而是要讓學生能夠將新知識與舊有的知識架構，統整後再重新加以建構。因此，在整個教學過程中，學生所有的迷思概念架構，都是改進教學的

重要資訊。

雖然教育的目的不是在於應付大小測驗，但不可諱言，測驗是學生學習的動機之一。本研究除彙整地球科學上常見的迷思概念外，也瀏覽許多地球科學的試題，找出哪些試題與這些常見的迷思概念有關，並進一步分析試題的設計是否能達到概念釐清的效果。希望這樣的探討，不單只是讓教師與學生知道哪些命題型式可以診斷迷失概念，更能讓他們在瞭解迷思概念所在後，可以進行知識重建而促成概念改變。

貳、迷思概念與改變

一、迷思概念的內涵

概念（concepts）一詞，從教育心理學的觀點來看，是針對一群事物的特定屬性，找出其共同性或相似性，以符號來代表之。此定義只強調概念符號所代表的意義，並沒有區分出概念的型態及其複雜層面。認知學派學者則視概念為一個心智活動，乃經由不斷的學習和經驗而獲得。經由不斷的學習與經驗，學生對於某一特定的概念的認知，逐漸由模糊進步到清楚，由具體簡單發展到抽象複雜，由純粹的對概念的理解逐漸發展到建立與其他相關的概念緊密的關係，進而應用此概念來解決問題。對於一概念的認知，也由於經驗的加深、加廣，漸漸形成一個可活用的認知結構。

耿正屏、陳瑞鴻、林素華、蔡顯（1991）指出概念的發展分為四個階段，分別為：（一）以原先具有的概念為基礎、（二）接受外來刺激、（三）內在的統整、（四）新概念的形成。概念的形成與學生學習的過程是很類似，學習者因為適應外在環境，改變原先概念或修正想法或基模，建構新的概念，使得能適應新的經驗。科學概念與日常概念都不是一次就完整形成，而是不斷演進的。

迷思概念（misconception）是指學生不正確的先前概念，迷思概念具有錯誤想法的涵義，並且強調學生接觸科學理論或模式之後，他們會結合某些不正確的訊息到他們的概念結構裡，最後又發展出不正確的想法。也就是學生對於自然現象，根據其日常生活經驗而發展出自己的想法，而這些想法中，有些是與科學家的概念有很大的不同，而這些不同科學家的概念就稱為迷思概念。迷思概念有以下特性：

（一）個人化及普遍化：個人的，又同時為多人所共享，具個別性及普遍性，但其與科學概念比較起來又不夠完備。

（二）頑強不易被改變：學生會排斥正式的科學概念，傾向以其既有的想法來解釋科學概念。或者先接受，但在一段時間後，先前所接受想法又有倒退的現象。

（三）跨越年齡、能力和國籍：不管在任何領域，任何國度，任何年齡的孩子均可能存在有迷思概念。

（四）不穩定性：個人認定，卻又前後不一致。有時同樣的問題，因其排列的位置，或是答題的順序，或所處情境的不同，學生常有許多不同的解釋，某些解釋容易出現，也容易拋棄。

（五）非隨機發生的：有些迷思具有歷史淵源，類似科學歷史的演進，今所說的迷思概念，可能是相同領域中早期人們所接受的想法。

二、概念的改變

邱美虹（2000）指出在有關概念改變的文獻中，學者大都指出科學概念學習困難的原因不外有下列幾點：（一）受到個人經驗的影響、（二）概念本身是抽象的、（三）概念本身是複雜的、（四）概念本身是微觀的。因此不同年齡、不同文化的孩童或成人在科學概念的學習上常易持有與科學家不同的另有想法。這些另有想法與科學家對科學現象的解釋不同，但受到上述因素的影響

而很難修正，例如：力學概念與學生日常生活的經驗有直接的相關，因此，即使已修過物理的大學生仍持有物體具有衝量的觀點。但另一方面，有些科學概念學生較易習得，如對孩童所進行的地球形狀的研究指出，孩童認為地球是平的或磁碟狀是不同文化中普遍存在的，但隨年齡的增長持有此觀點者逐漸減少。

Posner, Strike, Hewson and Gerzog（1982）指出如要讓學習者的概念有所改變，新概念則須具備以下四個條件：（一）使學習者對現有概念感到不滿、（二）新概念對學習者來說必須是能夠理解的、（三）新概念必須是合理的、（四）新概念必須是有效用的。而概念學習的結果，可能是內部知識的擴展，也有可能是內部概念結構做些許的調整，或者是概念結構的重組。

學生的知識認知結構，在已有的知識與新知識交互作用的過程中，會出現分化與統整的過程，此種現象稱為「概念的改變」。而學習就是舊知識與新知識在知識認知結構上的分化、重整、調適、成長與改變的一種過程。為了達成有效教學的目的，教師除了有必要在教學前先瞭解學生的先前想法，同時應針對學生的錯誤觀念來設計、編定教學內容。

邱美虹（2000）指出過去的研究對學生的迷思概念已有許多具體的資料顯示學生科學學習的困難，那在教學上我們可以如何做呢？過去學者針對迷思概念的教學策略提出許多作法，例如：先前知識與後設認知的培養，強調在課程規劃上不僅應考慮學科的架構，同時也應將學生想法一併列入考量，因而提出了以下幾點建議：（一）提供學生對自己思考反思的機會、（二）利用差異性事件使學生不滿意自己對現象的理解與解釋，進而進行概念改變、（三）蘇格拉底式的發問有助於學生警覺自己思考上不一致之處、（四）鼓勵學生進行有意義的建構以產生概念基模、（五）提供適當的學習情境以供學生了解科學的範疇與限制。

參、高中階段之地球科學迷思概念

國外地球科學迷思概念研究發現，孩童地球概念的發展隨年齡的增長由自我中心的觀念到部分自我中心，之後到科學觀念共容的觀念，影響孩童有關地球迷思概念的因素主要為「認知發展的層次」和「學校教育」。Philips（1991）將各年齡可能存在的地球科學相關迷思概念做一整理，發現各年齡對太空科學、岩石圈、氣圈、水圈和生物圈相關概念多少存在一些迷思概念，有些迷思概念存在於較為年長的人身上（例如：重力概念、太陽系的行星運行、宇宙、地球年齡、演化等），有些迷思概念存在於年幼孩童心中（例如：地球、地球形狀、月球與地球間的運行關係、晝夜、雲雨等），有些迷思概念是跨年齡以不同類型存在（例如：大氣特性、雨滴形狀、溫室效應、恐龍滅絕等）。

地球科學是一門與環境及人類生活密切相關的科學。其涵蓋的知識範圍領域廣大，除傳統上所區分的地質、海洋、大氣和天文外，環境變遷、地球資源與永續發展等，也可歸在地球科學內。地球科學除本身的領域專業知識外，更倚賴數學、物理、化學及生命科學等相關的專業知識。在自然界中，有許多現象是可以直接觀察到的，但是在這些可觀察的現象背後，其原理可能無法直接被觀察，因此，常常造成學生對於許多自然現象的迷思。

有關國內地球科學的迷思概念，陳淑筠（2002）參考歷年來的相關研究，彙整成相當完整的資料（表 1），並提出三點分析結果。

一、 根據表 1 顯示，學生在「天氣變化」、「組成地球的物質」、「地球運動」、「地球與太空」、「晝夜與四季」、「地表與地殼的變動」等單元主題上存有迷思概念。各單元主題的研究篇數，研究者也將之整理統計列於表上。從中我們可發現以「地球運動」、「天氣變化」及「地球與太空」的研究篇數及迷思類型最多，由此我們可推論，因為「太空」這個概念對於學生而言太過於抽象，因為教師並無法將「太空」與「地球」的全貌「拿」給學生看，他們接觸不到「太空」與「地球」的實體，只能藉由電影或

書本等媒介物來認識這些概念。所以學生在學習此主題單元時，就無法完全理解，而不求甚解的結果導致了學習上的困難。因此，教師們應致力於如何教導學生這些「抽象」知識，澄清他們的概念。

二、 天氣概念抽象難懂，但在日常生活中，我們天天在接觸的天氣現象，又如此熟悉。因此，學生常常根據自我的觀察與直覺的推論，對天氣現象加以解釋，而與實際的科學概念相差甚遠，因而形成了各種迷思概念。在國小自然科課程意見的調查中，氣候變化是國小教師及學生最感困難的單元之一（尚有觀測月亮、地球的運動、星球的運動等）。

三、 研究發現在日夜及四季成因主題中，由於概念較抽象，故學生多以直覺感官經驗來解釋。學生在學習時，因為原先對晝夜形成原因就不清楚，後來又受到四季形成原因的影響，就更加混淆了，於是呈現概念不穩定，顯然教師教太細節的部分時，忽略了學生對於整體概念是否了解。此外，在有關地球運動的概念上，由於地球運動屬於抽象概念，學生所具備的相關知識亦不足，所以學生在學習時易產生困難。故建議教師在教學時，不宜一開始就用討論方式教學，應先教授一些基本知識，且不要全用演講式教法，因為，本研究所分析的資料中許多相關研究都顯示，學生藉由彼此間的討論，可使其本來忽略或較模糊的概念，經由此的互動、刺激，而讓思考的層面增廣，澄清學生存有的迷思想法，再者，這也能使學生有多面向的思考，而不會侷限於線性方向思考。

表1　地球科學方面綜合研究一覽表（摘錄自陳淑蒨，2002，P56-60）

主題	篇數	綜合研究結果
天氣變化	5	綜合文獻發現學生存有下列迷思概念（魏金財，1992；許民陽，1996；洪志誠，2000；陳俊璋，2001；謝惠珠，2001）： ● 氣團為溫度相近的空氣聚集。 ● 氣團會改變氣溫的高低。 ● 氣團是颱風、鋒面、雲。

		● 氣團對天空的影響有帶來颱風、污染、寒流，以壞天空的想法為主。 ● 梅雨是氣團造成的結果（想法出現斷層）。 ● 梅雨發生在五、六月。 ● 梅雨是梅花開時下的雨。 ● 梅雨的形成原因是滯留鋒。 ● 梅雨是春雨、清明時節下的雨，在三、四月下雨，是梅花開花時下的雨。 ● 寒流來自大陸地區。 ● 寒流是滯留鋒造成的結果（想法出現斷層）。 ● 寒流來自太平洋。 ● 寒流的形成原因是冷氣團。 ● 形成寒流的原因是颱風、低氣壓、太平洋氣流。 ● 颱風的形成原因是暖濕的海洋空氣。 ● 颱風發生在五、六月。 ● 颱風是暖鋒、暖氣團造成的結果。 ● 天氣圖上的曲線是等壓線。 ● 該曲線的意義是氣壓的高低。 ● 濕度為空氣中的水份含量。 ● 天氣圖上的曲線是代表山的高度（等高線）、溫度高低（等溫線）、雨量多寡（等雨線）、乾濕程度。 ● 濕度代表下雨的多少、溫度的乾濕、除濕機的溫度、潮濕、乾燥。 ● 滯留鋒是屬於冷的特性。 ● 雲都是一整片。 ● 有藍色的雲。 ● 雲是黑色的。 ● 雲的位置沒有高低。 ● 雲不會增大或變小。 ● 雲由煙組成。 ● 不知道雲是由水滴組成。 ● 晴天的雲較厚。 ● 有雲就會下雨。 ● 雲變黑就會下雨。 ● 無雲之處也會下雨。 ● 低年級學童在晴、陰、雨的概念認知上，判別晴或陰的依據除太陽外，最重要者為冷熱及雲量多少。 ● 溫度和風也影響判斷，習慣上只要冷或風大就不

		是好天氣。 ● 判別陰或雨的主要依據為下雨與否及雲量。 ● 認為陰雨天看不見太陽時，是太陽沒升起或落下。 ● 不知道雲和下雨的關係。 ● 認為雲是一個地方或容器。 ● 認為下雨是水氣凝結而成。
組成地球的物質（岩石、水、大氣）	3	綜合文獻發現學生存有下列迷思概念（邱照麟，2000；郭純慧，2001；張凱綸，2001）： ● 多數學童對地球內部構造的了解仍以地表層為主，將可見之石材作為地球內部之主要成分。 ● 學童對「水氣態、液態的轉變」較具迷思概念。 ● 有關「可逆過程重量守恆」部分，學童具有物質狀態改變，重量也改變的迷思概念。 ● 有關「水狀態的轉變」部分，學童對水蒸氣及凝結存有較多迷思概念。 ● 有關「水循環」部分，學童對雲、霧及露的形成仍存有迷思概念。 ● 空氣逸出，致使餅乾變軟。 ● 空氣將餅乾的水份蒸發，致使餅乾變軟。 ● 高溫的空氣，致使餅乾變軟。 ● 杯中的水蒸發（水蒸氣），形成小水滴。 ● 杯內的冷空氣聚在杯外，形成小水滴。 ● 空氣未逸出，水也可以進入含空氣的容器中，未具「空氣佔有空間」的概念，認為垂直倒置於水槽中含空氣的玻璃杯，杯底的面紙會弄濕。 ● 非密閉容器，水可以阻擋空氣佔有容器空間，認為垂直倒置於水槽中含水玻璃杯，向杯內吹氣，杯中的水量不會減少。 ● 有的氣體有重量，有的氣體沒有重量，認為有些買來的氣球，可以飄上天空，是因為所裝的氣體沒有重量。 ● 風是不同氣流碰撞的結果。 ● 風是氣壓。 ● 風是集中在一起的空氣。 ● 風是水蒸氣變成的東西。 ● 風一直吹或風很大，所以風標會一直轉不停。 ● 風標箭頭的指向不是風吹來的方向。 ● 風大，氣壓會比較高，風小，氣壓會比較低。 ● 高的地方，氣壓會比較高，低的地方，氣壓會比較低。

		● 溫度高，氣壓會比較高，溫度低，氣壓會比較低。 ● 空氣的流動（風）會從氣壓高的地方流向氣壓低的地方，也會從氣壓低的地方流向氣壓高的地方。 ● 空氣的流動（風）是從熱脹處流到冷縮處。 ● 空氣的流動（風）會從熱脹處流到冷縮處，也會從冷縮處流到熱脹處。 ● 氣壓就是風。 ● 氣壓是風造成的。 ● 風是被氣壓推動的。 ● 在密閉的容器中，空氣的體積會隨著溫度的改變而改變。 ● 風大或風一直在吹，風標就會一直轉個不停。 ● 風標箭頭指向南，是吹北風。 ● 空氣受熱體積變大，重量也會變重，空氣受冷體積變小，重量也會變輕。 ● 空氣的組成中不包含固體雜質的成份。 ● 空氣中氧氣和二氧化碳含量最多。 ● 燃燒的煙會上飄與空氣的對流無關。 ● 煙會往上飄，因為煙和氫氣一樣，是比較輕的氣體。
地球運動	3	綜合文獻發現學生存有下列迷思概念（林秀鳳，1996；陳玉玲，2000；王景坤，2001）： ● 高空間能力者在地球形狀、重力概念、晝夜成因、四季成因的科學概念，皆優於低島間能力者。 ● 學生對方位的判斷不甚清楚，例如: 認為右手邊就是東方，左手邊就是西方，上面一定是北方，下面一定是南方等。 ● 地球自轉一圈需一年。 ● 地球每五分鐘轉一圈。 ● 地球不會自轉，只會公轉。 ● 我們人在地面上，地球是內部在轉，所以沒有感覺到地球在轉。 ● 人很小在地球的內側裡面，外側在轉，所以地球轉動我們沒有感覺。 ● 我們感覺不到地球在轉是因為有地心引力。 ● 我們走動時才會感覺到地球在轉。 ● 認階地球自轉的方向是順時針、自東向西轉，逆時針、自西向東轉。 ● 以地球的公轉，東、西方位是永遠固定不變的觀點，來解釋太陽的東升西落。

地球與太空	9	綜合文獻發現學生存有下列迷思概念（陳政瑜，1994；姜滿，1997；王景坤，2001；黃文美，2001；郭純慧，2001）： ● 學童認為，地球內部的溫度可能受到，太陽之有無、距地球中心之遠近、季節之變化… 等因素影響，而造成內部的有冷有熱。 ● 地球-月蝕模式: 地球遮蔽觀察者之視線或遮蔽月亮。 ● 月亮本體變形。 ● 太陽直射或雲遮蔽月亮。 ● 月亮是固定的。 ● 月亮的移動是以天數來計算的。 ● 地影遮住太陽，造成日蝕。 ● 月亮的形狀會改變主要是受到地球遮住月亮朦朧的地方。 ● 地球、月球轉動速度改變了，造成地影響月亮，所以形狀會改變。 ● 銀河系是許多星星聚集成一條銀色河流或織女和牛郎見面的地方。 ● 衛星是人造衛星或傳播資訊的工具。 ● 行星是會走動的星球。 ● 恆星是永恆的星星或有恆心的星星。 ● 流星可以許願、彗星會撞地球、彗星被稱為掃帚星。 ● 星星是角形ó（楊桃的橫切面），學生堅持親眼所見。 ● 有學生認階星星是＋形。 ● 學生認為星星比太陽近。 ● 有學生認為星星在月亮的前面。 ● 星星只是一個幻象。 ● 星星約是一個桌子大。 ● 星星和太空梭的大小差不多。 ● 固定在天空中，像一個躲避球大。 ● 星星像一間教室那麼大。 ● 星星反射太陽的光。 ● 星星反射月亮的光。 ● 星星反射太陽和月亮的光。 ● 星星是固定在天空中的。 ● 應該會移動，它會在原地繞小圈子。 ● 水星、金星檔住地球光線。

		● 烏雲遮到太陽。 ● 地球是天際間的星球與人類居住環境無關。 ● 地球是人類居住的球形星體，人住在地球中央的大平面，上、下半球皆為空氣。 ● 地球是人類居住的球形星體，人住在中央的平面上，上半球為天空，下為土地。 ● 地球是人類居住的球形星體，人類有的住在地球的外表面，有的居住在內部。 ● 地球是人類居住的球形星體，人類住在地球表面的各處。
晝夜與四季	1	文獻指出學生存有下列迷思概念（姜滿，1997）： ● 太陽白天出來，晚上躲到山後。 ● 雲或月亮遮住太陽。 ● 太陽繞地球每天一週。 ● 地球繞太陽公轉。 ● 地球每天自轉一週。 ● 雲擋住陽光，太陽強弱自行變化。 ● 太陽繞地球旋轉時，距離地球遠近不同。 ● 地球繞日公轉時，距離太陽不同。 ● 以公轉及地軸傾斜解釋四季變化。 ● 認為晝夜的形成，是因為地軸傾斜的緣故。 ● 地球受星球間引力的相互牽引而造成公轉。 ● 以地球的自轉來解釋四季的形成。 ● 以太陽直射赤道、南回歸線、北回歸線、或受光面積的大小，來判斷春、夏、秋、冬四季的位置。 ● 地軸傾斜或地球距離太陽的遠近造成了晝夜長短的變化。
地表與地殼的變動	1	文獻指出學生存有下列迷思概念（郭純慧，2001）： ● 學童認為地殼變動、板塊擠壓是地震的主要原因，但學童們對地殼如何會變動、板塊如何會擠壓並不是很清楚。 ● 關於地震發生的時機地點，學生認為地震的發生是無時無刻的。

雖然陳淑筠（2002）彙整出的地球科學迷思概念相當完整，不過其參考資料的研究對象大都是小學及國中階段的學生，隨著學生年齡的增加與科學知識的不斷累積，部分迷思概念可能已不存在於高中階段。因此，本研究另外參考

邱美虹、翁雪琴（1995）及許瑛玿、謝惠珠、鄒治華、張俊彥（2002）及基礎地球科學學科中心種子教師團隊（2011）的研究成果，彙整出幾個仍存在於高中階段的地球科學迷思概念，如下所列：

一、明亮的天空

學生只觀察到太陽的東昇西落與天空的明亮有直接關係，認為太陽升起了天空就明亮，太陽落下了天空就變暗。卻沒有瞭解到，天空明亮的背後是需要大氣層中空氣分子對陽光的散射作用。就像太空人在月球上，由於月球沒有大氣層，就算太陽高掛在天邊，天空還是一片漆黑。

二、四季成因

若問學生「為何冬天較冷，夏天比較熱？」許多學生可能會有錯誤的觀念認為：「因為冬天時，地球距離太陽遠；而夏天時，地球距離太陽較近所造成的。」或是，「地球面向太陽的那一邊是夏天，而背向太陽的那一邊是冬天。」

學生的迷思概念可能的原因：當靠近一個熱源時，容易被注意的是溫度變化，造成學生進而推論夏季的地球應該靠近太陽，相反的冬季遠離太陽。事實上，現在的地球除了自轉軸傾斜導致太陽直射區域不同，但夏季的地球比冬季的地球離太陽更遠，而夏季太陽直射北半球，冬季太陽直射南半球而導致季節差異。

三、上、下弦月

學生常有的錯誤概念：「上弦月與下弦月的區別是：月球的上半部亮，下半部缺時為上弦月，反之為下弦月」。事實是：（一）每個朔望月會出現 2 次「半個月亮」的月相。由朔到望（朔望月的上半期）之間出現的稱為「上弦月」；由望到朔（朔望月的下半期）之間出現的稱為「下弦月」（二）上弦月東昇西落之間，有時看起來是下半部亮，上半部缺，也就是說，上弦月的「上」字與

月球亮哪一邊毫無關聯，下弦月亦是如此。

四、視星等與亮度

亮度為星星看起來的明暗程度，為觀察者單位時間及單位面積所接受到星體輻射的能量。視星等（星等）為肉眼將星空中較亮到幾乎肉眼無法辨識之星分成 1 至 6 等，較暗者則為 7，8，9……，較亮者則為 0，-1，-2……。學生常誤以為數字愈高表示愈亮，因為這比較符合日常生活上的認知。

視星等的數字愈小，看起來愈亮，表示星星亮度也愈大，但星等和亮度並不相等，視星等為數字，有正值也有負值及 0，而亮度卻是一個物理量，包括了能量、面積與時間。星等數字愈小時為何會較亮？數字「-1」又怎麼會大於數字「1」呢？而且 0 等星亮度怎麼可能會是 1 等星亮度的約 2.5 倍呢？要解釋這問題只能說最早定星等的人當初就是如此定的，就像月考班上排名成績最高者（分數最多）卻是第“1”名，分數次高者反而是第“2”名，第一名分數高於第二名，但「1」卻是小於「2」的，排名較之星等所不同者是沒有第 0 名及負值。

五、引潮力

接觸潮汐現象，多數學生會死記「地球自轉一圈會經過 2 次滿潮和 2 次乾潮」，但常常只能說出地球面向月球一側，海水受到月球萬有引力的影響而發生滿潮，對於另一側的地球海水為何會是滿潮現象卻不能詳述。影響地球的潮汐水位高低的引潮力，主要受兩種力所影響，一為地月對偶旋轉離心力，另一為萬有引力。地球的引潮力即為此兩力平衡的結果。

引潮力與星球的質量成正比，但與其距離的三次方成反比，故太陽能質量雖遠較月球大，但引潮力卻比月球小。引潮力是學生常常誤解的概念，最困難的是為何地球兩端力量方向不一樣，甚至學生認為只有面對月球或太陽的那一面才有潮汐現象。

六、颱風

與颱風相關的迷思概念大都表現在「颱風為何種天氣系統」、「颱風結構」、「西南氣流引進豪雨」及「颱風、颶風與龍捲風的比較」等方面。學生常出現的颱風相關的另有概念有：（一）認為颱風是溫帶氣旋或寒冷的冷鋒、（二）認為颱風眼內的天氣狀況不佳（會有刮風及下雨的現象）、（三）認為風速最強的地方是在颱風外圍或颱風眼、（四）認為颱風引進西南氣流而帶來強風或是晴朗無雲的天氣、（五）混淆颱風、颶風和龍捲風。

颱風主要是發展於熱帶地區的低壓（氣旋）系統，由於發展於海洋，且熱量豐富，因此夾帶大量的水氣。而氣旋受到科氏力影響使得其結構更加完整，是一種劇烈的天氣系統,其中心是受下沉氣流影響,故颱風眼中心天氣是晴朗，但其颱風眼旁卻是雲系結構最為完整的區域。

七、氣壓

學生在氣壓的迷思概念有：（一）學生認為氣壓是風（氣流）所造成的，風越大氣壓越高、（二）同一水平面的兩個地區之氣壓與溫度關係，學生認為氣壓的變化與氣溫無關，或氣壓與氣溫成正比（氣壓高溫度高），或氣壓高低會使得溫度上升或下降、（三）在同一地點不同高度的氣壓變化發現，高度越高氣壓越高，學生認為高空空氣密度大於地面的空氣密度。

事實上，壓力是某一地區的空氣壓力（濃厚程度）所導致，但因表面受熱不均勻的差異導致具有高、低壓，或者與海拔高度所造成的差異。在同一地點區域壓力則會隨著高度越高，壓力減少；在同一水平面，因溫度高低不同造成壓力不均勻，如：高溫時容易造成該區域氣壓相較於周圍低。

李秀芬（1995）彙整會造成學生上述地球科學迷思概念的可能因素如下：

一、　學生易透過感官經驗形成迷思概念，因為學生運用其個人認知來詮釋自

然現象，其個人的直覺判斷而產生自我集中傾向之單一或片面的因果推理，也容易用不同理由來解釋相似的事物與現象。

二、 學生易受限於知識結構不佳情形下，使用籠統的、概括性認知來解釋所見的事物，並逐漸建構其概念，當科學名詞類似時，倘若未能仔細區辨、分化概念彼此間的差異，則會因誤解或混淆造成概念錯誤的連結。

三、 許多自然現象的迷思概念不易隨著年紀而改變，因個人認知在生活中具有其應用性，而概念僅有在部分情境下才出現侷限性，如考試情況下，否則是很難發現自己的迷思概念。

四、 若教師持有迷思概念，或教師發現學生的迷思概念未能有效的改變，迷思概念仍無法在學校教育中被去除。

肆、地球科學試題迷思概念探討

測驗（考試）是瞭解學生迷思概念的方法之一，能讓教師瞭解教學成效，進而調整教學方式，可以說測驗是教學的一部分。這裡所彙整的地球科學試題，與前面所提及高中階段仍存在的迷思概念有關，主要出處有：一、大考中心所命試題，包含歷屆學測、研究用試卷和學科知能量表等試題；二、高中教科書出版社之地球科學題庫，包含三民、全華、南一、泰宇、康熹和龍騰等；三、國中基測歷年試題。

由於大多數的試題只是測驗學生是否學會或記得課本上的知識，但並不是針對迷思概念的引導來設計試題，故以下所彙整出來的試題，雖然會測驗到與迷思概念有關的內容，但對學生消除這些迷思概念不一定有所幫助。因此，此研究會試著進一步分析試題所測驗的主要概念為何，題幹與選項的設計是否有助於學生理解迷思概念，或建議試題修改的方向等。

1. 測驗概念：明亮的天空

1-1. 圖 9 為美國太空人實地在月球白天進行表面觀察與拍攝的照片，證實在月球的白天，天空是黑暗的；但是地球的白天，天空是明亮的。依據同樣的道理，可以推論當太空人在水星、地球、火星等星球表面活動，在白天時比較其天空的明暗狀況，合理的是？（100 學測自然考科第 52 題）

圖 9

(A)水星比火星亮

(B)水星比地球亮

(C)火星比水星亮

(D)火星比地球亮

(E)水星、火星、地球會一樣亮

試題分析：本題是要測驗學生是否知道，明亮的天空是來自於空氣分子反射光線所致，而且空氣分子的多寡會影響天空的明亮程度。許多學生的迷思概念，認為只要是太陽有出現在地平線以上的話，天空看起來就是明亮的。

試題建議：本題的設計可以導正學生錯誤的概念，同時也測驗到學生的基本知識，是否知道水星、火星和地球的大氣狀況。但是，也有高中教師反應，九大行星的大氣特性是偏向純記憶的知識，且高中階段也不強調這些內容，建議在題幹上就提供學生這些訊息，只測驗學生有無天空明亮的迷思概念就好。

試題類型：開門見山，在題幹敘述上直接點出學生可能的迷思概念。

2. 測驗概念：四季成因

2-1. 若地球的自轉軸與繞日公轉軸的夾角由 23.5 度變成 0 度，則下列敘述何者正確？（84 研究用試卷自然考科第 6 題）

(A)沒有晝夜變化　　(B)每天日照時數不同

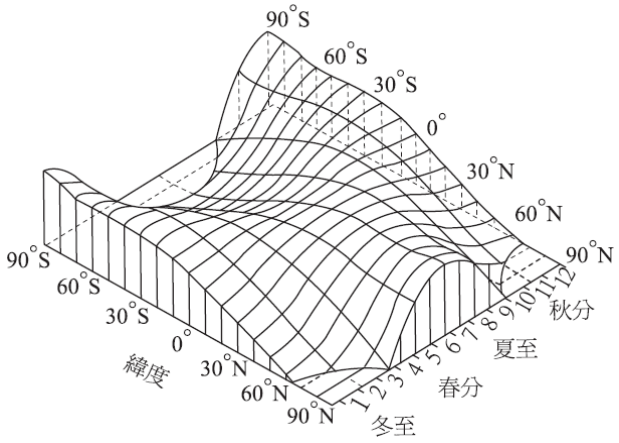
(C)沒有季節變化　　(D)一日等於一年

試題分析：本題要測驗地球自轉軸傾斜與否對各種自然現象的影響，而地軸傾斜是四季變化的主因。題幹中並未描述地球距離太陽遠近，刻意讓學生聚焦在地軸傾斜與四季的關係，對瞭解四季成因的迷思概念有所幫助。

試題建議：C 選項作為正確選項可能會有爭議，因為就算少了地軸的傾斜，地球公轉時距離太陽的遠近，也會對四季變化有所影響，只是影響的程度比地軸傾斜小。或許可將 C 選項修改成季節變化變得較不明顯。

試題類型：迂迴轉進，試題將迷思概念放在選項中。

2-2. 地球表面各緯度受到的太陽輻射強度隨季節的變化情形如下圖所示，請參考圖中資料回答 1～2 題。（本圖為立體模式，高度愈高代表太陽輻射愈強）（龍騰出版社地科題庫）



1. 以下敘述何者正確？

(A)夏至當天太陽直射北回歸線，臺灣地區的輻射強度大於兩極

(B)冬至當天太陽直射南回歸線，北極圈以北地區出現永夜現象

(C)一年當中，赤道受到的太陽輻射強度以夏至當天最強

(D)一年當中，赤道受到的太陽輻射總強度比兩極少。

2. 下列哪一緯度地區，太陽輻射強度的年變化最小？

(A)緯度 0 度　(B)緯度 23.5 度　(C)緯度 66.5 度　(D)緯度 90 度

試題分析：地球上的四季變化會受到太陽輻射強度的影響，不同緯度的四季變化情形，可由輻射強度的變化曲線略知一二，例如赤道地區四季變化不顯著，極區則相對極端。由於地球是一近似球體的星球，加上地球自轉軸傾斜 23.5 度，所以當太陽光平行入射到達地球時，太陽光會垂直入射的區域，隨著地球公轉，在南、北回歸線之間遊走。

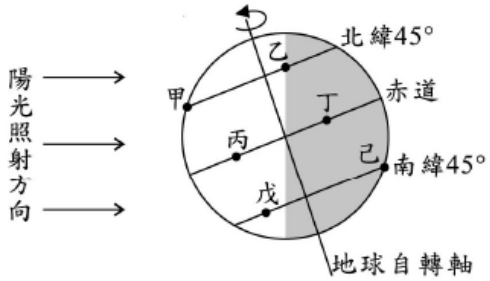
試題建議：除了太陽光的入射角度外，晝夜長短也會影響太陽輻射強度，所以在題幹描述上需要更完整一點。試題設計雖有創意，也讓學生注意到太陽輻射與四季變化的關係，但太陽輻射強度與太陽輻射均量定義是不同的，題幹描述上應更嚴謹一些比較好。

試題類型：開門見山，在題幹敘述上直接點出學生可能的迷思概念。

2-3.　請在閱讀下列敘述後，回答 54～56 題

圖(十八)為某時刻地球上晝夜分布示意圖，灰色部份表示夜晚區域，甲、

乙、丙、丁、戊、己為地球表面上六個不同地點。（97 第一次基測自然考科第 54~56 題）



圖(十八)

54. 哪些地點的白天比夜晚長？

(A)甲、乙　　(B)甲、丙　　(C)丙、己　　(D)丁、己

55. 若僅考慮太陽照射角度的影響，下列哪些地點的四季變化比較不明顯？

(A)甲、戊　　(B)乙、己　　(C)戊、己　　(D)丙、丁

56. 哪一個地點此時最接近正午時刻？

(A)甲　　(B)丙　　(C)戊　　(D)己

試題分析：此題組僅第 2 小題測驗學生對四季變化的瞭解，測驗學生是否知道四季變化的成因，與太陽光入射角度的變化及晝夜時間長短有關。

試題建議：試題僅詢問學生哪些地點四季變化比較不明顯，應可進一步詢問各地點的陽光照射角度變化情形，引導學生建立兩者的關連性，對迷思概念的改變會比較有幫助。

試題類型：迂迴轉進，試題將迷思概念放在選項中。

3. 測驗概念：上、下弦月

3-1. 在美國登月計畫中，阿波羅太空船上的太空人有許多機會從月球看地球，並且拍下畫面。圖 11 即為美國登月太空船中的太空人，在月球上空往地球方向所拍攝的影像，其前景(圖右下角部分)即為月球表面。在拍攝此圖的同一時刻，我們在地球上看月亮，看到的是哪一種月相？（97 學測自然考科第 36 題）

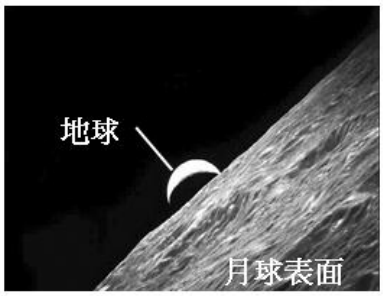
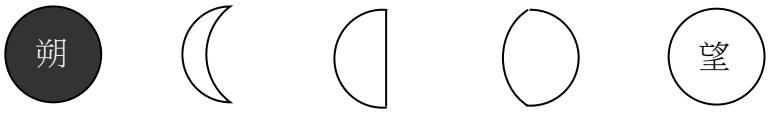


圖 11

(A) (B) (C) (D) (E)



試題分析：此試題的創意在於將觀測者放在月球上，再利用地球所呈現的模樣，推測日、月、地三者的相對位置，進而知道當時的月相。

試題建議：試題亦可設計成當月球是在上、下弦月的位置時，所見到的地球會是何種模樣，藉此加強學生的空間感，也有利於上下弦月的判別。

試題類型：隔靴搔癢，對釐清學生上、下弦月的迷思概念並無幫助。

3-2. 10-11 題為題組

「壬戌之秋，七月既望，蘇子與客泛舟遊於赤壁之下。清風徐來，水波不興。舉酒屬客，誦明月之詩，歌窈窕之章。少焉，月出於東山之上，徘徊於斗牛之間。」（摘自蘇東坡前赤壁賦）（90 學測自然考科第 10、11 題）

10. 下列何者最接近文中「壬戌之秋，七月既望」當晚之月相？

(A)  (B)  (C)  (D)  (E) 

11.「月出於東山之上，徘徊於斗牛之間」中之斗牛，是指什麼？

(A)二十八星宿中的斗宿與牛宿 (B)在東山之上的斗笠與水牛

(C)名為「斗」與「牛」的兩座高山

試題分析：第 10 題是與月相有關的試題，雖然以蘇東坡前赤壁賦來作為引言，但詩詞中的文字已經提到「望」，學生只要知道「望」就是滿月，然後從選項中找到此月相即可。

試題建議：月相變化是國中階段強調的概念，對高中生而言並無大太難度，建議以此概念為基礎往外延伸，以測驗多概念的整合能力。

試題類型：隔靴搔癢，對釐清學生上、下弦月的迷思概念並無幫助。

3-3. 發生於 2009 年 7 月的日食，讓居住於台灣地區的人有機會觀測到日食發生的經過。試問日食發生的當天晚上最可能觀測到下列哪種月相？（99 學測自然考科第 18 題）

(A)眉月　(B)弦月　(C)滿月(望)　(D)新月(朔)

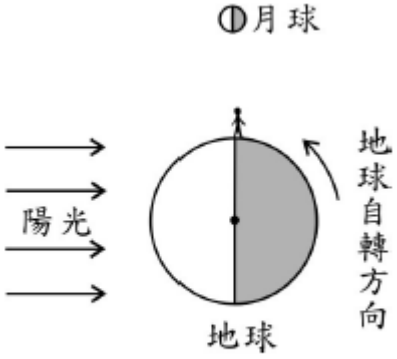



試題分析：此試題是測驗學生是否知道發生日食現象時，月相是新月（朔）。

試題建議：選項設計除了提供月亮圖片外，也有文字說明，學生在解答上難度降低許多，建議移除文字敘述。

試題類型：隔靴搔癢，對釐清學生上、下弦月的迷思概念並無幫助。

3-4. 圖（十六）為陽光照射地球示意圖，此時有關人所在位置的時間及月相的敘述，下列何者正確？（95 第二次基測自然考科第 43 題）



圖(十六)

(A)此時為中午，月相為上弦月　(B)此時為黃昏，月相為下弦月

(C)此時為子夜，月相為上弦月　(D)此時為清晨，月相為下弦月

試題分析：此試題需判斷兩個概念，一是地球的晝夜區分，二是月相。學生可藉由地球自轉方向來推斷月球的公轉方向，知道月相正經歷由望到朔的過程，故此時呈現的月相是下弦月。

試題建議：試題可延伸測驗學生月球的公轉情形，因為月球繞地球公轉是月相變化的成因，可藉此更清楚學生對此概念的瞭解程度。

試題類型：迂迴轉進，試題將迷思概念放在選項中。

4. 測驗概念：視星等與亮度

4-1. 已知甲星與乙星都是造父變星，而且甲星的視星等為 9，乙星的視星等為 14。測得兩星的亮度變化週期相同，表示甲、乙兩星的光度應該相同，則乙星的距離是甲星的幾倍？（提示：$M = m + 5 - 5 \log d$，其中 M 為絕對星等，m 為視星等，d 為距離；絕對星等與光度密切相關）（88 學測自然考科第 19 題）

(A)0.1 倍　(B)1 倍　(C)10 倍　(D)100 倍　(E)1000 倍

試題分析：試題已提供計算式，學生須瞭解視星等與絕對星等的定義，以及距離對視星等的影響，然後將數字帶入計算式即可。

試題建議：試題設計可強化學生星等與數字間的連結，建議可延伸出另一試題，測驗學生肉眼觀測情形，以消除可能的錯誤概念。

試題類型：開門見山，在題幹敘述上直接點出學生可能的迷思。

4-2. 表 2 為恆星資料，依據表中恆星的顏色、視星等、絕對星等，判斷哪一顆星與地球的距離大於 32.6 光年？（絕對星等為星星距地球 32.6 光年的亮度）（99 學測自然考科第 12 題）

表 2

恆星	顏色	視星等	絕對星等
甲	黃	4	6
乙	藍	8	11
丙	紅	7	9
丁	白	3	2

(A)甲恆星　　(B)乙恆星　　(C)丙恆星　　(D)丁恆星

試題分析：試題中恆星顏色的設計是陷阱，因為顏色與距離無關。表中只有丁恆星的視星等數字大於絕對星等，而在星等中的數字愈大表示看起來愈黯淡。

試題建議：試題設計很有創意，也讓學生有機會注意星等與數字的關連，有助於釐清學生這方面的迷思概念。

試題類型：迂迴轉進，試題將迷思概念放在選項中。

4-3. 夜間觀賞星空，會發現每顆星星各有各的亮度，也有不同的顏色，天文學上用視星等將不同的星星亮度分級，而絕對星等是指把星星放在指定距離時，星星所呈現的視星等。表一將一些恆星的視星等、絕對星等與表面顏色分別列出來：（97 研究用試卷自然考科第 45 題）

表一

星　名	視星等	絕對星等	表面顏色
天狼星	-1.46	1.4	白色
五車二	0.08	-0.5	黃色
北極星	1.97	-3.64	黃白色
角宿一	0.98	-3.55	藍白色
心宿二	0.96	-5.28	紅色
大角星	-0.04	-0.31	橙色

試運用表一中資料判斷恆星表面溫度以及恆星與地球的距離，並選出下

列敘述，哪些正確?（應選三項）

(A)天狼星與地球距離最近　　(B)北極星與地球距離最遠

(C)心宿二與地球距離最遠　　(D)五車二表面溫度最低

(E)角宿一表面溫度最高

試題分析：題幹中已經描述視星等與絕對星等的定義，試題是要測驗學生是否知道星等與數字間的關係，還有恆星表面顏色所代表的溫度高低。

試題建議：視星等的數字小於絕對星等，表示該恆星距離地球小於 32.6 光年，反之則大於 32.6 光年。試題藉由問學生恆星距離的遠近，強調星等與數字的關連，藉此釐清學生此概念是很好的切入點。而恆星表面顏色則與星等和距離無關，只表示溫度的高度，所以這是測驗另一概念。

試題類型：迂迴轉進，試題將迷思概念放在選項中。

4-4. 下表為 5 顆恆星的相關資料，請根據資料回答下列問題。（龍騰出版社地科題庫）

（距離模數：m－M=5logd－5；m：視星等；M：絕對星等；d：與地球之距離）

恆星	甲	乙	丙	丁	戊
視星等	3.5	0.1	0.8	1.4	1.0
絕對星等	5.7	4.5	2.2	－0.3	－3.4
顏色	白	橙	黃	藍	紅

1. 哪顆星的發光強度最大？　(A)甲　(B)乙　(C)丙　(D)丁　(E)戊

2. 哪顆星離我們最遠？　(A)甲　(B)乙　(C)丙　(D)丁　(E)戊

試題分析：此試題在測驗學生絕對星等與發光強度的關係，以及星等與距離的關係，這兩個小題都可以加深學生對星等與數字的連結。

試題建議：表中的恆星顏色在問題中並未提到，感覺是多餘的資料，但也有可能是設計來混淆學生，亦是不錯的作法。

試題類型：迂迴轉進，試題將迷思概念放在選項中。

5. 測驗概念：引潮力

5-1. 海面週期性的升降現象，稱為潮汐；滿潮時與低潮時的水位差稱為潮差。下列有關潮汐的敍述，何者正確？(應選兩項)（94 研究用試卷自然考科第 41 題）

(A)全世界各地區的潮差均無太大差異

(B)朔、望日時的潮差大於上、下弦日

(C)每日滿潮或低潮的時刻較前一日延遲約50分鐘

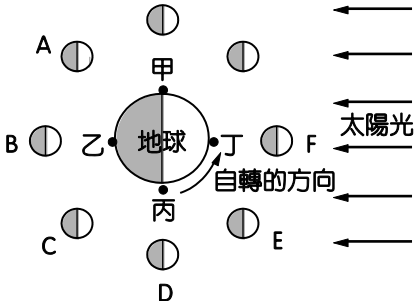
(D)潮汐發生的原因是受到天體萬有引力的影響，其中以太陽的引潮力影響最大，月球其次

試題分析：題幹已經描述潮汐現象及潮差定義，主要是測驗學生幾個與潮汐相關的常見問題，例如：各地潮差、大小潮時的月相、漲退潮的延遲、太陽和月球的影響大小。

試題建議：每個選項都可以獨立成一個試題，這樣試題的設計可以更廣更深，而且只測驗一概念，對學生釐清問題會更有幫助。

試題類型：隔靴搔癢，對釐清學生引潮力的迷思概念並無幫助。

5-2. 下圖為在北半球上空俯瞰地球、太陽、月球的相對位置圖，若不考慮海流、地形等因素，僅單就日月引潮分布來看，小潮當天何時發生滿潮？（翰林出版社地科題庫）



(A)中午 12 時、子夜 12 時　(B)清晨 6 時、傍晚 6 時

(C)上午 9 時、夜晚 9 時　　(D)不一定

承上題，今日發生大潮經過數日後出現小潮，表示月球在哪兩點間移動？

(A)A → B　　(B)A → C

(C)B → D　　(D)B → F

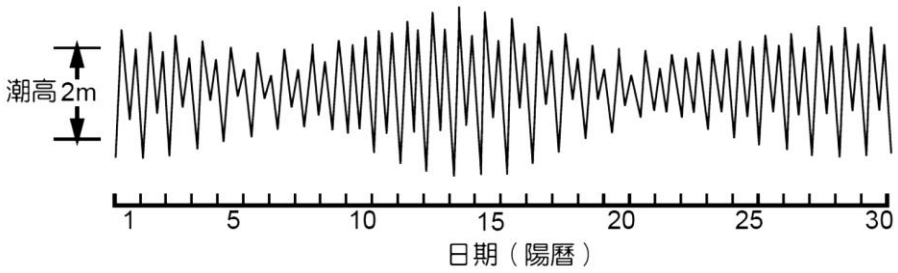
(E)D → E　　(F)D → F

試題分析：試題除了測驗學生是否知道大、小潮時，地球、太陽和月球的相對位置外，還要能從圖上辨別白天與晚上的時刻，觀測者的所在時刻反而是學生比較不容易建立的概念。

試題建議：試題結合大小潮與漲退潮時刻是頗有難度的設計，此一概念應可延伸出更多樣的測驗方式。

試題類型：隔靴搔癢，對釐清學生引潮力的迷思概念並無幫助。

5-3.　下圖表示 1989 年 11 月 1 日至 11 月 30 日止，這段期間某一個港口的潮位變化曲線，下列敘述何者錯誤？（全華出版社地科題庫）



(A)造成每日水位最高及最低各有兩次的原因，是地球自轉之故

(B)在11月14日的月相為新月或滿月

(C)在一個月內該港口的最大潮差約4公尺

(D)每月滿潮與乾潮的差不一定，但都以兩週的週期作變化

(E)在 11 月 6 日時，日、月、地在一直線上

試題分析：學生必須從附圖找出一些資訊，例如：一天內有兩次漲退潮，一個月內有大潮和小潮，及潮差大小等。除此之外，學生也必須了解這些潮汐的變化與地球自轉和月球公轉有關，更要能進一步判斷月相。

試題建議：試題設計將多個概念放在一起，答題上並不容易，也無法明確判斷學生是哪一個概念不清楚，所以一個試題測驗 1~2 個概念較為理想。

試題類型：隔靴搔癢，對釐清學生引潮力的迷思概念並無幫助。

6. 測驗概念：颱風

6-1. 28-29 為題組

圖 6 是 2005 年龍王颱風自 9 月 30 日 12：00 到 10 月 3 日 0：00 的颱風路徑圖，圖上所標示的時間為台灣地區時間(月/日)， 每個標示點間

隔為 6 小時。根據圖 6 的資料， 回答 28-29 題。（96 學測自然考科第 28-29 題）

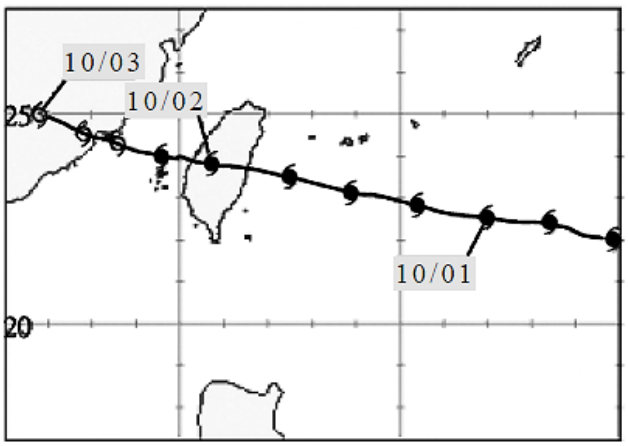
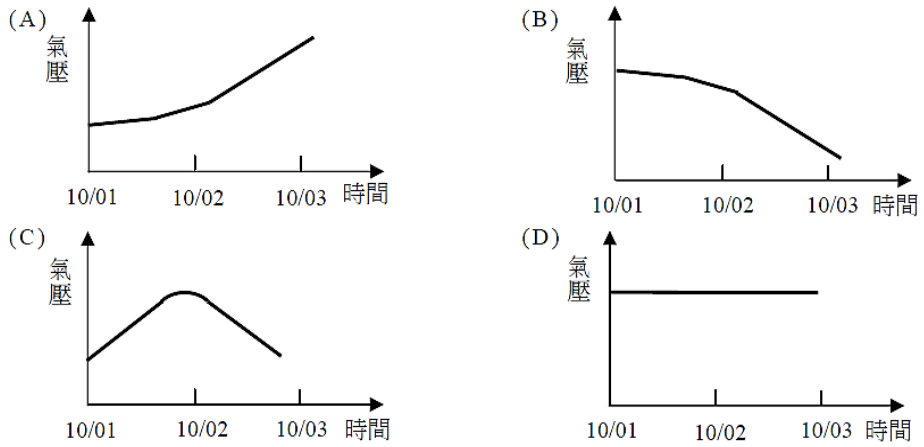


圖 6

28. 下列哪一圖最能代表颱風中心氣壓自 10 月 1 日到 10 月 3 日的變化？



29. 有關龍王颱風的敘述，下列哪一項正確？

(A)生成於花蓮東方100公里的海面上

(B)發生在9月、10月，容易引進西南季風

(C)朝東北轉向後減弱

(D)容易造成台灣東北部地區發生豪雨

試題分析：第 28 題的設計除了需要學生知道颱風是低氣壓系統外，並從附圖瞭解到颱風隨時間變強或變弱，再推論到颱風中心氣壓的變化，頗有挑戰性。而第 29 題則是測驗學生颱風的特性與天然災害，比較偏向記憶性。

試題建議：第 28 題可進一步詢問學生，造成颱風中心氣壓有此種變化趨勢的原因為何，以強化學生這方面的概念。

試題類型：迂迴轉進，試題將迷思概念放在選項中。

6-2. 2009 年 8 月莫拉克颱風侵台，造成八八水災。使得台灣地區重大的人員傷亡， 重挫台灣地區的經濟和農業。圖 4 為莫拉克颱風路徑圖， 路徑圖中標示為甲、乙、丙、丁的哪一段時間，最可能為台灣地區帶來豪雨？（99 學測自然考科第 21 題）

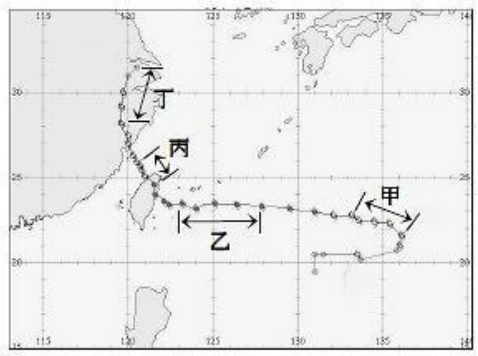


圖 4

(A)甲　　(B)乙　　(C)丙　　(D)丁

試題分析：莫拉克颱風所帶來的豪雨與西南氣流的引進有關，因此學生必須從附圖判別颱風位於何處，才有機會引進西南氣流。

試題建議：試題設計讓學生可簡單作答，只要選何時颱風最靠近臺灣即可，雖然這樣的推論不完全正確，但卻可答對，造成試題的鑑別有誤。建議可測驗學生臺灣哪些地區降下豪雨，與颱風和西南氣流得關連性。

試題類型：隔靴搔癢，對釐清學生颱風的迷思概念並無幫助。

6-3. 51-52 題為題組

圖 17 是象神颱風 89 年 10 月 31 日 20 時的紅外線衛星雲圖， 試根據此圖回答 51-52 題。（90 研究用試卷自然考科第 51-52 題）

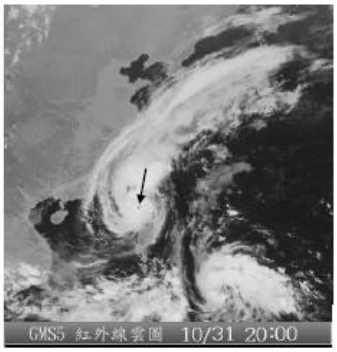


圖 17

51. 圖17中箭頭所指颱風中央之圓形黑點內的天氣狀況應為何？

(A)強風驟雨 (B)綿綿細雨 (C)烏雲密佈 (D)晴朗無雲

52. 下列有關象神颱風的敘述何者錯誤？

(A)象神颱風為逆時針旋轉之氣旋

(B)圖中箭頭所指颱風中央之圓形黑點， 應為颱風眼的位置

(C)颱風風速最大處位於颱風之暴風圈邊緣處

(D)89 年 10 月 31 日 20 時台灣全島應皆有下雨

試題分析：題幹利用天氣雲圖讓學生知道颱風眼沒有雲分布，然後在第 51 題希望學生藉此推知颱風眼的天氣狀況會是晴朗無雲，而且其他選項的安排也有強調此一特性，有助於學生破除此迷思概念。而第 52 題的問答涵蓋比較多概念，例如氣旋的旋轉方向、颱風眼位置及最大風速等，學生需概念清楚才行。

試題建議：第 52 題可針對颱風最大風速來設計試題，例如在附圖上標示 4 個位置，分別位於颱風眼、眼牆、暴風圈內、颱風外圍，雖然試題難度不高，但主要是可協助學生釐清迷思概念。

試題類型：開門見山，在題幹敘述上直接點出學生可能的迷思概念。

7. 測驗概念：氣壓

7-1. 冬季時假設北京和高雄的地面氣壓相同，但是北京的地面溫度遠比高雄的地面溫度低，則下列哪一敘述不正確？（101 學測自然考科第 15 題）

(A)北京的飽和水氣壓比高雄的飽和水氣壓低

(B)北京的近地面空氣密度比高雄的近地面空氣密度大

(C)北京與高雄兩地單位面積上空的空氣重量大約相同

(D)北京的地面露點溫度一般比高雄的地面露點溫度低

(E)近地面處北京的氣壓隨高度下降的變化比高雄慢

試題分析：此試題測驗多重概念，學生需瞭解氣壓、空氣密度、空氣重量和溫度之間的關連性，並進一步延伸至水氣壓和露點溫度的比較，是屬於偏難的試題。

試題建議：雖然可達到提醒學生注意相關概念，但若試題設計只針對氣壓、空氣密度和溫度的話，對學生這方面的迷思概念會更有幫助。

試題類型：迂迴轉進，試題將迷思概念放在選項中。

7-2. 下列有關氣壓隨高度變化的敘述，何者正確？（87 學測自然考科第 23 題）

(A)氣壓隨高度之變化率為每1公里約升高6.5百帕(毫巴)

(B)氣壓隨高度之變化率為每10公尺約降低1百帕(毫巴)

(C)氣壓在垂直方向的變化比它在水平方向的變化小

(D)空氣愈緻密時，氣壓隨高度之遞減率愈小

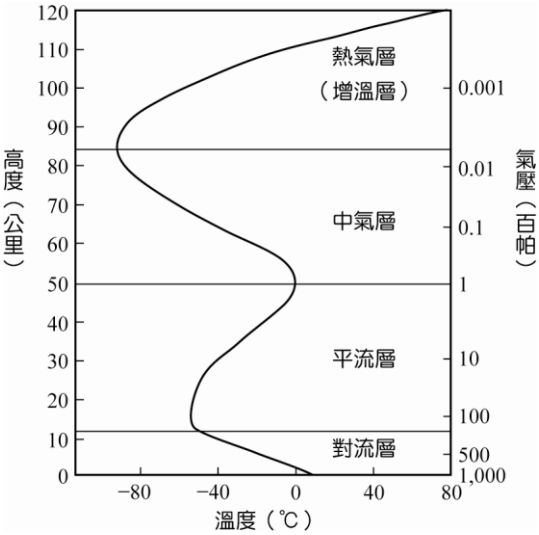
(E)空氣愈稀薄時，氣壓隨高度之遞減率愈大

試題分析：學生在氣壓方面的迷思概念，有高度愈高氣壓愈高或高空空氣密度大於地面空氣密度，在這個試題選項裡都可以得到釐清。

試題建議：D 和 E 選項的設計並不理想，有強調氣壓是隨高度遞減的提示。或許可詢問學生空氣的疏密程度是否會影響氣壓隨高度的增減率。

試題類型：迂迴轉進，試題將迷思概念放在選項中。

7-3. 大氣溫度及壓力隨高度的變化情形如下圖所示，在高度 80 公里以下的大氣層中，下列哪些性質通常會隨高度而減少？（應選三項）（康熹出版社地科題庫）



(A)水氣含量　　(B)空氣密度　　(C)空氣溫度

(D)空氣壓力　　(E)氧氣所占的體積百分比

試題分析：從附圖上的資訊，學生應該知道大氣壓力會隨著高度遞減，且就大尺度來看，大氣壓力與溫度變化是沒有關連的。試題是要學生聯想，當大氣壓力隨高度遞減時，還有哪些大氣性質也會減少，對學生在氣壓隨高度變化的迷思概念有所幫助。

試題建議：此試題若單考學生空氣壓力隨高度的變化則太容易，因此設計成多選題是不錯的作法，而且選項設計上有很高的相關性，可有效鑑別學生能力。

試題類型：開門見山，在題幹敘述上直接點出學生可能的迷思概念。

7-4. 藤光到海水浴場玩，發現白天與夜間風吹的方向剛好相反。下列有關此現象的推論，何者正確？（94 第二次基測自然考科第 7 題）

(A)白天時吹海風是因為海水面的溫度較高，而陸地上溫度較低

(B)夜間時吹陸風是因為海水面的溫度較低，而陸地上溫度較高

(C)白天時吹海風、夜間時吹陸風是因為海水與陸地比熱不同

(D)白天時吹海風、夜間時吹陸風是因為海水與陸地密度不同

試題分析：海陸風的形成，是來自於海面及地表溫度不同，對空氣所造成的高、低壓力差所造成的。溫度高的區域會形成低氣壓，溫度低的區域會形成高氣壓，這是學生容易有迷失概念的地方，錯以為溫度高的區域是高氣壓。

試題建議：海水和陸地的比熱雖然扮演重要角色，但試題加入這個選項反而會模糊焦點，若以改變學生迷思概念為前提的話，選項應略做調整。例如以何時吹陸風何時吹海風，以及海水面和陸地的溫度高低，交叉設計選項即可。

試題類型：隔靴搔癢，對釐清學生氣壓的迷思概念並無幫助。

綜觀以上試題，是否協助學生更瞭解相關的迷思概念，簡單分類如下：

一、 開門見山：

這類試題在題幹描述或圖表中，或多或少就點出了迷思概念，藉此先讓學生建立好正確概念後，才以此概念為基礎，進一步測驗學生的相關知識或推理能力，例如：1-1、2-2、4-1、6-3、7-3。這類試題在輔助學生釐清迷思概念上，效果應該會比較顯著。

二、 迂迴轉進：

這類試題將迷思概念放在選項中，若學生在此概念已有迷思的話，答題難度上自然提高不少，或容易被其他選項所誘答，例如：2-1、2-3、3-4、4-2、4-3、4-4、6-1、7-1、7-2。雖然學生可能無法順利作答，但試題設計至少能達到提醒學生注意這類迷思概念的效果。

三、 隔靴搔癢：

這類試題並未觸及迷思概念的核心，感覺只是在外圍打轉，對學生瞭解迷思概念沒有幫助。但這不是試題本身的問題，而是應該反過來檢討，教師或許已經知道學生有這樣的迷思概念了，但卻未針對此迷思概念來命製試題，例如：上、下弦月（3-1~3-3）及引潮力（5-1~5-3）的試題。這樣的情形值得我們注意！

伍、結論與展望

迷思概念的確認有其嚴謹性，例如陳彥任（2007）採用二段式診斷測驗來確認大氣迷思概念。本研究是倚賴許多學者先進的研究成果，以及高中教師的第一線經驗，才得以彙整出部分成果。而概念改變本身是一個長期且緩慢的過程，國內外學者提出許多理論與方法，例如許瑛玿、謝惠珠（2004）所採用的BDEI（Conceptual Bridging – Differentiation – Exchange - Integration）概念改變教學策略，所以不能期待只依賴測驗就能順利達到概念改變的效果。

目前的教學體制，在課堂上一個教師需面對 30 位學生左右，每位學生因知識背景的差異，對同一概念可能表現出不同的迷思，造成教學上的困擾。但若要一一去導正學生的迷思概念，所要投入的時間和精力是難以想像的。另外，雖是不同的求學階段，但試題測驗目前仍扮演著重要角色，在此一現象未改變前，若能在試題設計上多下點功夫，無論是採開門見山或迂迴轉進的方式，不斷在試題中融入這些學生容易產生的迷思概念。除了可達到提醒學生的目的外，

甚至可以慢慢產生迷思概念上的釐清與調整，是此研究的主要目的。

地球科學係為一種嘗試以合於邏輯的推論，來解釋生活週遭及大自然所發生的現象，並以此建構對客觀世界認知的一門學問。因此，老師在幫助學生學習地球科學的時候，不能只強迫學生「記住知識」，因為今日的真理，明日未必依然為真，而是要幫助學生學習到「思考的方式」，進而能建構出自己的「思考模式」，來發現問題、處理問題與解決問題。當學生產生迷思的時候，或許正是一個引導學生重新思考的好機會。一個聰明的老師應該要好好把握這時機，在學生面臨知識認知衝突的時候，引導其建立邏輯的思考方式，從而建立其正確的知識認知架構。

參考文獻

李秀芬（1995）。**高中學生氣壓概念另有架構之研究**（未出版之碩士論文）。國立成功大學，臺南。

邱美虹、翁雪琴（1995）。國三學生「四季成因」之心智模式與推論歷程之探討。**科學教育學刊**，**3**（1），23-68。

邱美虹（2000）。概念改變研究的省思與啟示。**科學教育學刊**，**8**（1），1-33。

耿正屏、陳瑞鴻、林素華、蔡顯（1991）。**我國國中學生生生物概念發展生物體內物質的運輸**。行政院國家科學委員會專題研究成果報告（編號：NSC80-0111-S-018-03-D），未出版。

基礎地科學科中心種子教師團隊（2011）。**專有名詞研發**。未出版。

許瑛玿、謝惠珠、鄒治華、張俊彥（2002）。調查台灣地區國中學生颱風概念理解現況。**科學教育月刊**，**255**，2-11。

許瑛玿、謝惠珠（2004）。應用概念改變教學策略在颱風常識的學習。**師大學報**，**49**（1），15-40。

郭重吉（1988）。從認知觀點探討自然科學的學習。**教育學院學報**，**13**，351-378。

陳彥任（2007）。**中學生「二段式大氣迷思概念診斷測驗」的發展與應用**（未出版之碩士論文）。私立中原大學，桃園。

陳淑筠（2002）。**國內學生自然科學迷思概念研究之後設研究**（未出版之碩士論文）。國立臺東師範學院，臺東。

Philips, W. C. (1991). Earth science misconceptions. *The Science Teacher, 58*(2), 21-23.

Posner, G. J., Strike, K. A., Hewson, P. W., & Gertzog, W. A. (1982). Accommodation of a scientific conception: Toward a theory of conceptual change. *Science Education, 66*, 211-277.

大陸高考外語學科聽力考試的發展歷程和命題工作

劉慶思

大陸教育部考試中心

摘要

聽力是外語學科的重點考查內容之一，但因其實施操作難度較大，長期以來一直被各方人士視為畏途。由本文所介紹高考外語聽力考試的發展歷程和命題工作可以看出，為全面考查考生的語言交際能力，教育部考試中心在聽力考試的研究、推廣和試題命製方面投入了巨大的力量，也取得了令人滿意的成果。

關鍵詞：高考、PETS、外語、聽力

劉慶思，大陸教育部考試中心命題三處處長

Listening Test of the Matriculation English Test in Mainland China

Qing-Si Liu
National Education Examinations Authority, PRC

Abstract

Listening is one of the key components in foreign language testing. However, because of the difficulty in administration, it is regarded complicated and dangerous in undertaking and thus always put on the shelf. As can be seen from the history of implementing the listening test and the test development affairs, the National Education Examinations Authority (NEEA) values the assessment of the candidates' communicative language ability and makes listening test a necessary part in the matriculation foreign language test after a great effort.

Keywords: University Entrance Examination, PETS, English Test, Listening

---

Qing-Si Liu, Director of Foreign Languages Tests Department, National Education Examinations Authority

壹、引言

外語是高考若干學科之一，長期以來一直與語文、數學一起列為所有考生的必考科目，受到考生的廣泛重視。外語學科多年來一直提供英、日、俄、德、法、西六個語種的試卷供考生自由選擇；各個語種均提供含聽力和不含聽力兩類試卷，由各省份根據自己的特定需要進行統一選擇。此外，各語種均含筆試和口試兩種考試形式。筆試為所有考生之必考，考查考生的聽力、閱讀理解和寫作能力，以及對語言知識的掌握情況；口試則專門為外語類考生（報考外語專業和國際金融、國際貿易等專業的考生）所設計，考查考生的外語口語交際能力。外語學科在高考中所發揮的作用及外語學科命題的複雜性，由此可見一斑。瞭解下文所介紹聽力考試的發展歷程和命題工作所考慮各事項後，大家會更為真切地感受到：聽力考試的推進和命題工作殊為艱難。

貳、聽力考試的發展歷程

聽力是語言交際能力的必要組成部分，是語言教學中一項不可或缺的重要內容，因此，自然應該在語言測試中得以體現。然而，由於實施條件的限制和組織管理的複雜性，聽力考試一直都是在克服各方面困難的情況下艱難推進。大致來講，教育部考試中心的高考外語聽力考試經歷了以下幾個發展時期：

一、聽力考試研究期（1993 年至 1999 年）

《全日制普通高級中學「英語教學大綱」（初審稿）》1993 年推出，1996 年開始由全國各高級中學統一使用。使用該「大綱」的學生于 1999 年高中畢業，參加全國普通高校招生統一考試。與以往「大綱」不同的是，該「大綱」注重學生語言交際能力的培養，特別是在聽和說兩方面提出了明確的要求；同時，在「考試、考查」部分，該「大綱」明確要求，「測試的形式要包括筆試

和口試或聽力測試」。人民教育出版社出版的英語教材，著力體現了對學生聽力和口頭表達能力培養的要求，同時也帶動了中學英語師資力量的加強和語音設施的改善。

為了配合該「大綱」的使用，教育部考試中心 1993 年起即著手進行大規模考試中增加聽力部分可行性的研究。首先，在湖北省進行了「高中會考中增加聽力部分」的研究，該項研究 1993 至 1996 連續進行了三年，證實了會考中加設聽力部分的可行性，隨後，其他一些省市也陸續在會考中增加了聽力部分，這在聽力設施、考務經驗等方面為各省在高考中增加聽力部分奠定了基礎。1997 年起，開始在廣東省進行高考英語中增加聽力部分的可行性研究，亦取得了可喜的成果，證實了高考中增加聽力部分的可行性。1997 年在廣東的高考英語試卷中，僅將聽力部分作為參考分處理；1998 年按 10%的權重計入英語科總分，1999 年則按理想全重 20%計入英語科總分。

二、聽力考試推廣期（2000 年至 2004 年）

考慮到「教學大綱」對聽力教學的明確要求，基於高考英語增加聽力部分的可行性研究成果，2000 年，教育部決定在高考英語試卷中逐步增加聽力部分，並針對各省、自治區、直轄市中學師資、辦學條件發展不平衡的情況，制訂了過渡方案：2000 年，教育部考試中心向各省提供不含聽力的高考英語試卷、聽力部分占全卷權重 13%的試卷和聽力部分占全卷權重 20%的試卷，供各省選擇。2001 和 2002 年，僅向各省提供含聽力部分的高考英語試卷，某些條件不成熟的省可自行去掉聽力考查內容，對其餘試題的分數進行加權處理。2003 年，如無特殊原因，各省均須採用含聽力部分的高考英語試卷。這樣，高考英語科開始了以增加聽力考查內容為重點的考試內容和形式的調整。

三、聽力考試反思期（2005 年至 2006 年）

作為高利害考試，高考中出現的任何問題都會引起社會的廣泛關注，影響

到社會穩定。外語試卷中增加聽力部分以來，每年的考試都會因各種原因出現不同的問題，如：聽力設施故障、磁帶問題、考務工作人員失誤等，這在一定程度上影響了考試的安全性、穩定性和權威性。為了切實解決這些問題，教育部在 2005 年制訂的《教育部關於做好 2005 年普通高等學校招生工作的通知》中提出：「從 2005 年開始，普通高校招生對考生外語聽力測試不再做全國統一要求。各省級教育行政部門可根據本地教育實際自行決定外語聽力測試的考試形式、時間和計分辦法，並將成績在錄取時提供給有關高等學校。」根據這一精神，教育部考試中心徵詢各省的意見後，決定此後每年向全國提供含聽力和不含聽力的外語試卷各一套，由各省進行選擇。結果是：個別省開始重新採用不含聽力部分的試卷。

四、聽力考試完善期（2007 年—現今）

（一）課程標準基礎上的高考

從 2004 年起，大陸開始進行高中課程改革試驗，以新的課程標準為教學依據。與以往的教學大綱相比，課程標準在課程目標、課程理念、教學內容、教學方式等方面都有較大的變革。外語學科課程標準與原大綱相比也發生了很大的變化，主要為：提高了對中學生的要求，強調語言運用能力的培養，明確了對各項語言技能的要求。在此基礎上設計的高考外語試卷自然應該對考生的聽力水準進行檢測，以對中學外語教學產生正向的反撥作用。為此，教育部考試中心經過多方調研設計出了課程標準基礎上的含聽力英語試卷，並於 2007 年首次使用，開始了聽力考試的再次推廣期。按照課程標準在大陸推廣的進度，預計 2015 年各省均會採用這一類型的高考英語試卷。

（二）高考英語聽力使用 PETS-2 級聽力部分考試

PETS-2 級考試是 Public English Test System（PETS）中的第二級，考查內容和試卷難度與高考英語科大致相等，每年舉行兩次。

中學進行課程改革後，各省考試機構逐步開始使用含聽力的外語試卷，但仍視聽力考試的組織實施為畏途，希望能夠採取措施適當降低該考試的風險，於是開始考慮採用 PETS-2 級考試中的聽力考試。

根據相關省份的要求，PETS-2 級為考生單獨設置了聽力考試，考生在高考前有兩次參加的機會，考試機構將選擇兩次考試中的較高成績計入考生的外語學科總分。這就既給予考生多次參加考試的機會，又有效降低了聽力考試的利害程度，該舉措受到了中學師生的歡迎。在這些省份，高考期間舉行的外語考試理所當然地去掉了聽力部分。

參、聽力考試的命題流程

聽力考試的命題工作與試卷中的其他部分同時進行，但相比而言需要更多的步驟，操作起來也更為複雜。為使大家全面瞭解聽力部分的命題工作，下面將分別介紹採用題庫形式和入闈形式的聽力命題流程。

一、PETS-2 級中聽力考試的命題工作

為了保持試卷難度穩定，應對一年多次考試，PETS 考試自問世以來一直採取以試題為單位的題庫式命題方式，這種情況下聽力部分的命題工作主要可分為以下幾個階段：

（一）徵題

在每年的特定時間，學科秘書都會根據題庫維護結果和當年的工作計畫，以書面形式向命題教師分派徵題任務，內容清楚、任務詳盡是任務書撰寫所遵循的基本原則。命題教師需根據「命題教師指導手冊」和任務書的各項要求，保質保量地完成任務，並按時郵寄給學科秘書。

（二）試題預編輯

收到命題教師寄來的試題後，學科秘書按照規則將這些試題歸類整理，剔除明顯不符合要求的試題。如有必要，還需進行簡單的格式處理，使每位教師的試題除編碼不同外，其他均保持一致。

（三）試題編輯

試題編輯以會議的形式進行，具體工作任務為對命題教師交付的試題逐一進行審核、加工，並標注上各種定性的屬性，如：語篇類型、話題、語域、試題考點等，以備存入題庫之用。

（四）試測組卷

為獲取所需資料和檢查試題品質，任何一道進入題庫的試題都必須經過試測，因此，試題編輯會結束後，學科秘書即馬上著手將所命制的各個試題，按照特定的組卷要求，如：大致難易度、考點分佈等，組配進嵌入錨題的幾套試測試卷中。

（五）試測試卷錄音

聽力是考試中非常重要的一部分，實施該部分必不可少的一項工作是試卷錄音。為了順利完成各年度大量的錄音工作，教育部考試中心與某專業的音像部門以協議的形式，確立了長期、順暢的合作關係，考試中心負責錄音內容的把握和安全監控，而音像部門負責有償提供音棚和技術人員。

（六）試題試測

試題試測計畫的制訂和組織管理是一項複雜而艱苦的工作。得知所需試測試題的數量後，試測管理部門即需著手試測人數的計畫和試測點的聯絡、確定工作；為確保試題安全，試測時則需派專人赴試測點，負責過程管理、安全監

控、試卷銷毀等各項工作。

（七）合格試題入庫

試測完成後，統計人員將利用經典和專案反應模型分別對每套試測試卷進行分析，確定每個小題的易度（facility）、區分度（discrimination）、難度（difficulty）等數值。學科秘書對各小題仔細研究，剔除一些資料不夠理想的試題後，將合格試題及定性和定量的題目屬性一起鍵入到題庫中。

（八）正式考試試卷組配

每次正式考試前，學科秘書需根據特定要求完成正式試卷的組卷工作。在此過程中，學科秘書關注的焦點為：語言材料話題的均衡、試題考查點和難度的分佈、詞彙的總數量等。

（九）正式試卷的審核和校對

試卷組配完成後，學科秘書將邀請命題組長和另一位骨幹教師對其進行審核和校對，主要工作任務是審查試卷的難度和試題的搭配情況，查找可能存在的科學性和格式體例方面的錯誤。

（十）正式試卷錄音編輯

考慮到考生的感情，正式考試的錄音工作往往在試測錄音的基礎上重新進行，這一是可保證試音與正式考試音質、音效的一致，減輕考生的心理壓力；二是可以保證各段錄音材料在錄音效果方面的一致性，有助於保證錄音品質。

（十一）考後統計與試題參數入庫

考試實施後，考務部門將對考生的答卷情況進行抽樣分析，並將基於正式考試的統計資料回饋給命題部門，學科秘書則會將這些實測資料錄入題庫，補充試題資料方面的屬性。

（十二）已啟用試題的處理

已啟用的試題將會被轉入「退休庫」中，進入若干年的休眠期。之後，仍有可能出現在不同類型的英語考試中。

二、高考外語科聽力部分的命題

為確保考試安全，高考外語科的命題工作近年來一直採用入闈方式，要求命題人員在特定的時間內（一個月左右），命制出含聽力部分的若干套完整試卷。因此，聽力部分的命題工作隨特定試卷一起進行，省卻了題庫式命題中的入庫、試卷組配等環節，這種情況下的命題工作主要包括如下環節：

（一）徵題

徵題工作的內容和要求與 PETS-2 級基本一致。

（二）試題預編輯

命題教師在會議地點集中後，學科秘書收齊老師們所帶的 U 盤和紙面稿，將試題和相關原始材料拷貝到命題專用電腦上，並按相應規則進行編碼和格式處理；同時會剔除掉某些明顯不符合要求的試題，以節省編輯會的時間。

（三）試題編輯

試題編輯為入闈命題的主要工作內容，具體工作任務為對命題教師交付的試題逐一進行審核、加工，並組配成若干套試卷；工作方式為，學科組分成若干小組，對試題進行挑選、加工，並在適當的時間進行交換、審核，以進一步提高試卷品質。

（四）錄音編輯與光碟製作

聽力考試所特需的錄音工作與 PETS-2 級考試一致。

肆、聽力考試的考查內容

聽力部分要求考生能夠聽懂有關日常生活中所熟悉話題的簡短獨白和對話。具體來講，考生應能夠：

一、理解主旨和要義

任何一段對話或獨白都圍繞一個中心內容展開，從對獨白或對話主旨的總結情況可以看出考生對整個語篇的宏觀掌控能力，因此，每年的試卷中都有以各種形式出現的這類試題。

二、獲取事實性的具體資訊

具體資訊是指語篇中的時間、地點、人物等細節性資訊，是說話者傳輸主旨要義，表達其意圖、觀點所需借助的必要資訊，在聽力考查中佔有重要地位，每年均占試題總量的 70%左右。通過考查對這些資訊的掌握情況，可以看出考生對語篇中關鍵資訊的識記、分辨和把握能力。

三、對所聽內容做出簡單推斷

話題與語言情景、談話者的身份、談話人之間的關係等有著密切的聯繫，所以，由所談論的內容即可以對談話情景、說話人的身份等進行推斷，每年的試卷中都會有若干這類試題。對這些問題的解答可以反映一個人對口頭英語的理解程度，因而成為聽力測試的重點考查內容之一。

四、理解說話者的意圖、觀點和態度

說話者總會有說話意圖，或是交流資訊，或是探討問題，或是闡述自己的想法，表明自己的態度或意見，而這些意圖有些是明說出來的，有些則隱含在話語的字裡行間，需要聽者自己去揣摩、推斷。對這一能力的考查，可以看出考生對口頭英語的分析理解能力。

伍、聽力錄音的基本規範

高考英語聽力的錄音由專業人員在專業的錄音棚中完成，因此，錄音中無任何噪音，答題時間控制極為準確。此外，高考英語科的錄音還遵循了以下規範：

一、英語播音人主要來自英國，使用的是標準的倫敦英語。有時，播音人可能來自其他以英語為母語的國家，但其標誌性的發音得到了一定程度的控制，單詞的發音基本為英國英語。

二、播音人的語速比一般英美人正常談話的速度稍慢一些；另外，語速隨語篇的不同而有所變化。當對話為家庭成員間、同事間的談話時，語速會稍快一些；當對話雙方為陌生人時，語速則稍慢一些；而獨白的語速則隨話題和聽眾的不同而有所變化。

三、為營造語言情景，對話和獨白前可能會出現一些必要的背景音，如電話鈴聲、敲門聲、汽車發動或剎車聲等。這些背景音有助於考生瞭解對話情景，理解對話雙方的關係，正確解答試題提出的問題。

陸、聽力考試語言材料的選取

語言材料選取在聽力測試中扮演著非常重要的作用，合適的材料可以幫助全面、準確地對考生的聽力進行考查，加強考試的區分作用，有助於提高考試的信度和效度。高考外語試卷設計人員在為聽力部分選材時著重注意了以下幾項原則。

一、聽力材料的多樣性

每年的聽力部分均由兩節構成，第一節中設置的為基於短對話的單個試題，第二節則為基於長對話或獨白的多個試題，因此，試題設計即對語言材料

在長短、形式等方面提出了多樣性的要求。此外，高考外語試題在選材方面還堅持了以下多樣性原則：

（一）談話情景的多樣性

談話情景是指談話發生的場合、地點或背景，如朋友聚會、餐館聚餐、機場告別、機場接人、接電話等。任何對話都發生在特定的情景之中，而不同情景中所使用的語言會有一些不同，有時如離開該特定情景可能就難以理解，因此，對對話情景的推斷成了聽力所考查的語言微技能之一。為了充分考查考生對不同情景中英語的掌握情況，瞭解考生對必須的英美文化的瞭解程度，每年高考英語科的聽力部分都會設計多個不同的語言情景，另外還會專門設計若干推斷語言情景的試題，如：2011 年試卷中設計了 10 個不同的情景。

（二）談話人之間關係的多樣性

社會上的每個人都具有多重身份，如：一個男人可能的身份包括兒子、丈夫、父親、同事、顧客、領導、下屬等，談話時他會根據自己不同的身份和會話場合選擇不同的詞語，採用不同的語氣，因此，由這些內容可以推斷談話人之間的關係。另外，因話題與談話人有著密切的聯繫，所以，由談論的話題也可以推斷談話者之間的關係。每年外語試卷的聽力部分都通過提供不同類型的語篇，不同場景下發生的對話向考生展示說話者之間多種多樣的複雜關係。

（三）談論話題的多樣性

總的來看，高考外語科聽力材料的談論話題均可歸入日常生活一類，但這一話題又可以進行細分，如：購物、餐飲、交通、健身、假期生活等。每年試卷中的十個語篇均會涵蓋多個話題，如:2011 年英語試卷的聽力部分涉及了 8 個不同的話題。話題的多樣性可以說明拓寬考查內容的覆蓋面，從不同的角度檢測考生的聽力水準，確保考查結果的準確性。

（四）材料類型的多樣性

材料可以是獨白，也可以是對話。獨白可能為新聞播報、天氣預報、課堂中授課的部分內容，也可能為對自己或其他人一段往事的回憶；談話則可能是在任何場合，就任何話題進行的對話。材料類型的多樣性同樣有助於拓寬考查內容的覆蓋面。

二、語言材料中沒有任何生詞

生詞量在外語學習者中是一個較為敏感的話題，考生在試卷中看到會覺得很彆扭，其他人在考後看到則會在不同的場合進行渲染，似乎試卷中出現了某一類型的科學性錯誤。從語言學習的角度看，這應該說是極不正常，因為在實際的語言交際中，任何人都可能會遇到自己尚未聽到或看到過的詞語，而語言情景可以在一定程度上幫助聽者或讀者正確地理解這些資訊；再者，根據語言情景和上下文中的內容猜測或推斷未知的資訊也是語言能力的一個重要方面。即使如此，為了照顧考生的情緒，高考外語科在聽力部分還是暫時採取了消除一切生詞的做法，因此，聽力部分無論是錄音內容，還是試題中都沒有超出「考試說明」詞彙表的單詞。當然，合成詞和派生詞不在生詞之列。

三、文章結構完整，內容豐富

聽力部分中的各個語篇，無論篇幅長短、內容多少，基本都能夠構成一個獨立的整體，從話語結構上看有頭有尾，有起始有發展，沒有任何語篇給人以非常唐突，不知來龍去脈的感覺；從話語內容上看，則資訊豐富、語境清楚、對話者身份明確。

四、語言材料和試題難易度符合考生的實際水準

多年的外語各語種試卷說明，聽力材料的話題一般限於日常生活，無生僻內容；語篇長度基本都在 300 詞以內，不會過於加重考生的記憶負擔；考生所

需要的背景知識均在普通高中畢業生所應瞭解和掌握的英美文化常識之內；另外，語篇中沒有生詞，說話者語速稍慢。這些都使得語言材料的難度沒有超出考生的實際語言水準。

五、盡力保留所選材料形式與內容的真實性

多年來試卷聽力部分的錄音表明，高考外語各語種非常重視所選擇聽力材料語言的真實性，即形式和內容的真實性。形式真實性是指每一段錄音聽起來都是口述內容，有明顯的口語特點，諸如猶豫、停頓、重述等；內容的真實性指的是，錄音內容主要來自英美國家的原版材料，雖然有些材料經過了一定程度的改寫，但在語境和言語方面符合英美國家日常生活的實際情況。

柒、聽力考試試題的命製

錄音所需的聽力材料確定後，試題設計人員即著手挖掘材料中的各類資訊，進入試題命制階段。該階段的工作要求他們不僅有較高的外語水準，掌握一定的語言測量理論，還要瞭解中學的聽力教學情況。在日常的具體工作中，命題人員主要關注了以下事項：

一、把握語言材料中的關鍵資訊

對語言材料的仔細分析是命題階段的首項工作，試題設計人員需要理清材料中的各類資訊，包括整個語篇的主旨要義和支撐性資訊、各個語段中的總體資訊和細節資訊、不同說話者的觀點態度、說話者之間的關係、談話的場合等。每個語篇中都有大量的這類資訊，不可能均反映在試題中，所以，命題時把握關鍵資訊至關重要。此處的關鍵資訊是指一些總體資訊（主旨要義、觀點態度）和重要但容易理解失誤的細節資訊。

二、合理推斷考生理解失誤情況

為聽力部分命制試題時，在構思出正確答案的同時，必須為每個小題設計出兩個干擾項（錯誤選項）。干擾項的設計，要求試題設計人員清楚考生外語學習中的難點和弱點以及在理解不同語篇時可能遇到的特殊問題。

三、試題表述語言簡練

過多的閱讀量會影響對聽力內容的理解，同時造成難以分析考生理解失誤的原因，所以，儘量減少試題閱讀量成為命制各類考試聽力試題的原則之一，高考外語各語種試卷的聽力部分亦然。分析多年的高考試卷後可以發現，聽力部分各小題的題幹均用詞較少，非常簡練，選擇項用詞也同樣節儉。以 2011 年英語試卷為例，20 個小題的題幹總用詞量為 157 個，平均每小題僅 7.85 個；20 個小題中，只有 4 個小題的選擇項採用句子形式，其它均為單詞或短語；另外，20 個小題的 60 個選擇項共使用了 207 個單詞，平均每個選擇項用詞僅為 3.45 個。

四、試題分布均衡

聽力試題分佈的均衡性體現在三個方面：首先，在帶有兩個以上小題的語篇中，重要資訊處均設置了相關試題，且各小題按照錄音內容的順序排列，相隔一定的時間，保證了試題在語篇中的平均分佈，使考生有時間在新的資訊到來之前完成上一個問題。其次，該部分的 20 個小題一般都會覆蓋聽力所要求的各項語言微技能，如：2011 年的英語試卷中有 1 個小題考查觀點態度，兩個小題考查主旨要義，5 個小題考查推斷能力，12 個小題考查對具體資訊的把握能力。最後，試題分布的均衡性還體現在難易試題的分佈上，作為選拔性考試的部分考查內容，聽力部分的試題通過難、中、易試題的合理搭配，充分發揮了對不同層次考生的區分作用。

捌、結語

對於外語考試而言，聽力部分的增加具有里程碑式的意義，標誌著外語科考試內容的日趨完善，會對外語教學產生強烈的正向反撥作用。但新事物在推出過程中難免會出現一些問題，遇到一些阻力，考試機構一定要為各種可能出現的事件制訂出相應的應急預案，做好充分的思想和物質準備。

參考文獻

中華人民共和國教育部（1993）。**全日制普通高級中學英語教學大綱（初審稿）**。北京：人民教育出版社

中華人民共和國教育部（2003）。**普通高中英語課程標準（實驗）**。北京：人民教育出版社

教育部考試中心（2010）。**全國英語等級考試考試大綱（二級）**。北京：高等教育出版社

教育部考試中心（2012）。**高考文科試題分析（2012 年版）**。北京：高等教育出版社

雲海工程：高考分數產生、報告、解釋和使用的改革努力

韓寧

大陸教育部考試中心

摘要

「雲海工程」是大陸教育部考試中心和雲南、海南兩省的考試機構共同進行的與高考分數產生、報告、闡釋、使用相關的一籃子改革試點。該專案的目的是根據《國家中長期教育改革和發展規劃綱要》建設專業考試機構的要求，學習借鑒國際教育考試領域的最新理念和技術，挖掘高考資料中蘊藏的豐富資訊，向考生、中學、大學、教育行政部門、命題部門提供更加豐富的資訊回饋服務，從而完善高考的技術環節，提高高考的服務水準。其中，面向考生的分數報告和分析服務改變了傳統考試「一張紙條幾個數位」的分數報告方式。利用資料倉庫和資料採擷技術，考生的學習成就可以從更多的維度和更多的細節被提供出來，這為高考探索建立「綜合評價、多元錄取」的新機制提供了技術條件。面向中學和教育行政部門的資訊回饋服務則著眼於考試的後果和影響，發揮高考對基礎教育的反撥作用，為教育行政部門提供決策參考，幫助中學改進教學。本文分析了國際考試行業標準對分數產生、報告、闡釋、使用的要求和一些世界著名考試項目的技術現狀，介紹了「雲海工程」改進報名環節、探索建立量表分數、探索實現年度間分數可比性和省際分數可比性、探索提供診斷性分數報告、探索建立標準參照分數解釋、探索實現「增值評價」的技術任務，和在這些技術基礎上建立高考資訊回饋和服務網站的總體思路。通過過去兩年的試點，大陸考試中心和省考試機構建立了新的業務運轉模式，形成了規範的高考「初始分析」和「最終分析」的業務流程。

關鍵詞：雲海工程、分數產生、分數報告、分數闡釋、分數使用、常模、標準、等值

韓寧，大陸教育部考試中心評價研發中心主任

Project YunHai: An Effort on Generation, Peport, Interpretation, and Utilizing of GaoKao Scores

Ning Han

National Education Examinations Authority, PRC

Abstract

Project YunHai is a joint reform effort on score generation, report, interpretation, and utilizing of GaoKao in Yunnan and Hainan by NEEA together with the provincial examination boards of the two provinces. The purpose of the project was to improve the technical processes and upgrade the quality of service of GaoKao by providing more information feedback service to candidates, high schools, universities and colleges, and government departments. The service of online score report and analysis to candidates has changed traditional way of score report which was described as "several figures on a piece of paper". By technique of data warehousing and data mining, more detailed and multi-dimensional achievement levels of candidates can be obtained, which made the goal of "comprehensive evaluation, admission by multiple channels" technical feasible. By focusing on the impact and consequence of GaoKao, the information feedback service to high schools and government bodies helped the decision making of government bodies and the improvement of teaching of high schools. The requirements on score generation, report, interpretation, utilizing by authoritative international standards and the best practices from some well known exam programs were introduced and analyzed. The details of Project YunHai were introduced which included improvement of registration, research on scaled score, research on comparability of scores across different years and regions, research on diagnostic score report, research on yielding criterion referenced score interpretation, research on value added assessment, etc. An information feedback and service web site has been set up based on above research. After two years' pilot study, a new business model between NEEA and provincial exam boards and a two phase standardized data process flow including initial analysis and final analysis have emerged.

Keywords：Project YunHai, Score Generation, Score Report, Score Interpretation, Score Utilizing, Norm, Standards, Equating

Ning Han, Director, Center for Assessment Research and Development, National Education Examinations Authority

大陸高等學校入學統一考試（高考）最重要的作用和目標當然是為國選才（材），這絕不會引起任何爭議。但是，選拔從來不是高考的唯一目標。長期以來，高考的作用都被說成是「兩個有利」甚或「三個有利（助）」。「兩個有利」即「既有利於高等學校選拔合格新生，又有利於引導中學教學改革」，此種提法多見於 1977 年恢復高考之後到上個世紀末，如原國家教委考試中心在上世紀八九十年代出版的考試說明中對高考的原則基本都是這樣說的。2001 年頒發的《基礎教育課程改革綱要》（http://www.edu.cn/20010926/3002911.shtml）指出：高考的基本原則是「有助於高等學校選拔人才、有助於中學實施素質教育、有助於擴大高等學校辦學自主權。」這裡的前兩條和原來意思相似，只是用 更有時代特徵的「素質教育」代替了指向簡單的「中學教學」。2010 年頒發的《國家中長期教育改革與發展規劃綱要（2010-2020）》（http://www.gov.cn/jrzg/2010-07/29/content_1667143.htm）是未來 10 年國家教育發展的綱領性檔。它將高考的基本原則重新界定為「有利於科學選拔人才、促進學生健康發展、維護社會公平。」應該說，較以往站得更高，更加全面了。但是，這卻賦予了高考更加艱巨的目標，將高考之外的很多社會矛盾都集中到了高考身上，幾乎提出了一個「不可能完成的任務」。從選才自身角度考慮，如何進行選拔才更加科學、合理和公平是永恆的爭論話題，很難形成亙古不變的結論。如目前在大多數省區市普遍採用的平行志願（http://baike.baidu.com/view/288569.htm），由於滿足了社會追求公平的心理受到了普遍歡迎，但從人才選拔的角度看，平行志願把考試分數的作用推到了極致，除了總分高低，考生其他一切特徵統統視而不見，造成了高校錄取分數的「扁平化」，誇大了考試分數之間的細小差別，加劇了高校生源大戰。這很自然引起人們對「綜合評價、多元錄取」（見《國家中長期教育改革與發展規劃綱要（2010-2020）》）的要求。然而，在競爭日益激烈、社會誠信缺失的總體

環境下，「綜合評價」卻更容易引出人們對「暗箱操作」的擔憂，於是又呼喚「裸考」，這就形成了一個怪圈。如果從高考自身跳出來，從教育和社會發展的更高層次來看，高考改革固然要呼應群眾的呼聲，關注「誰能上大學」和「誰不能上大學」，更應著眼於考試的後果和影響，引導人才選拔觀念的變化，引導基礎教育的發展，引導千百萬學生的成長與成才之路。這樣，才能最大限度地發揮高考在選拔之外更多的功能。某種程度上說，這也是高考對帶來諸如「應試教育」等副作用的自贖之路。否則，廢科舉前車之鑒不遠。

即使在大多數從事此項工作的專業人士眼中，考試的核心環節無外乎命題和考務。尤其是目前考試舞弊出現集團化和高科技化、考試安全環境惡劣的條件下，考試機構的大部分精力被用在了「保安全」上，幾乎是無暇他顧。考試機構的業務流程往往自命題始，然後通過試卷印刷、發送、保管等物流環節，基層工作人員安排考場、組織考生參加考試，試卷回收和評卷。評卷之後的事情就簡單了，經過一系列簡單的加法，每個考生被告知獲得的總分和各個科目的得分，這幾個分數（尤其是總分）就成為了招生的絕大部分依據—如果不說成是全部依據的話—很大程度決定了每個考生的命運，同時也在決定著教師、學校、教育當局的政績和獎懲。

客觀地說，動用了大量公共資源組織、深刻影響了億萬學子人生軌跡的高考的主要功能就是為了對每一個考生作出能否被錄取進心儀大學的重要決定。難怪俗語有「分、分、分，學生的命根!」其實，分數也是家長、老師乃至更多的人的「命根」。在很多地方，它甚至已經成為政府對教師、學校、教育管理當局如何獎懲的依據。如此高利害高風險的推論和判斷，有理由要求更高的可信度和有效性。

大規模社會化考試在西方發達國家已成為一種專門化技術和行業，成為一種和教師在課堂上用來檢查學生掌握情況從而改進教學的課堂測驗完全不同的社會公共服務。提供這種公共服務的既有非營利機構，也有上市公司，翹楚

者如美國的 ETS、ACT，英國的劍橋評價（Cambridge Assessment）等領導世界學術風氣。在這些專業化考試機構中，考試的流程已完全不能僅用命題和考務兩個環節來概括，而是在命題和評卷之外，至少還應該包括一個專業性更強、技術更複雜的重要環節：即考試結果的產生、報告、解釋和使用。而這個環節，正是我國大規模社會化考試與世界最先進水準的最大差距所在！國內有人將這一環節稱作「評價」，並未被廣泛接受。國外在學術分類上通常歸入「心理計量學」（Pasychometrics）。

客觀地說，就命題環節而言，我們聘請的老師和專家是在用製作藝術品的態度來命題，考試機構通常還要用「入闈」這種在西方人眼裡已很不近人道的方式來保證題目的安全和準確，拋開文化差別不談，從整體上我們的題目品質不應輸於人。在考試的硬體設施上，經過三十年的改革開放，很多地方甚至可以令洋人「羨慕嫉妒恨」了，考試中心觸角遍及全國的視頻指揮系統就曾多次讓來自西方考試機構的參觀者矯舌不下。中國悠久的考試歷史和傳統更應該使我們在考試組織方面信心滿滿，但是，在考試分數的產生、報告、解釋和使用上，實事求是地說，我們確實差距甚大。這正是我們急起直追的背景和原因。

壹、國際權威考試標準對考試分數的產生、報告、解釋和使用的規定

由美國教育學會 AERA、美國心理學會 APA、美國教育測量學會 NCME 共同制定的《教育和心理測驗標準》（AERA，APA，NCME，1999）是教育考試和心理測驗領域跨行業的「聖經」，規定了美國心理測驗和教育考試行業的入門門檻。該標準的主要內容分成三部分，分別是考試的研發、質量評估和文件建設，考試中的公平性，考試的應用。該標準使用了大量篇幅對考試分數的產生、報告、解釋和使用做出了明確的規定。

關於分數產生，該標準的第四章專門以「量表、常模和分數的可比性」為題，首先對與考試分數有關的核心概念如常模、分數可比性、等值、常模參照和標準參照、標定、臨界分數等做了解釋和界定，然後通過 21 條具體標準條款提出了每個技術環節應該達到的要求。這些條款的絕大多數都是針對所謂匯出分數或量表分數的，但也沒有排除原始分數的使用。該章第一條（4.1 款）就強調，任何考試都應該建立完備的文檔和手冊，將分數的具體含義、合理闡釋和局限性告知考試分數的使用者。21 條條款涉及常模的建立過程、分數等值的技術細節、臨界分數的確定辦法，體現了建立在科學測量和評價理論上的現代考試和傳統的經驗型考試的根本區別，橫跨心理學、教育學、統計學、資訊技術等諸學科，是現代教育測量理論體系的技術核心。

關於分數報告，第 5.10 款說「當發佈考試資訊時，考試機構應該負責提供關於考試分數的正確解釋」，考試機構提供的資訊應該包括考試的內容範圍、分數的通俗含義、分數的精確性、常見的誤解、合理的用途等。第 13.14 款說「分數報告應該伴隨著對每個分數點或分數等級的測量誤差範圍的清楚的描述，同時還要提供關於如何解釋分數的資訊」。

關於分數闡釋，核心的兩個概念是標準參照的分數闡釋和常模參照的分數闡釋。特別應該注意的是該標準在提到常模參照和標準參照時所使用的術語是 Norm-referenced score interpretation 和 Criterion-referenced score interpretation，這就是說，標準沒有把它們看做兩類考試，而是看作兩種不同的分數闡釋方式，注意到這一點是非常有意思的，等下筆者還會講到。技術上說，和常模參照分數闡釋相伴隨的往往是「常模參照組」；和標準參照分數解釋相伴隨的往往是「臨界分數」，這可以說是區分兩種分數闡釋方式的一個標記。

關於分數使用，該標準特別注意公平性，將公平性視作考試的生命和基石，但該標準也清醒地認識到，「公平」一詞從來就沒有準確的定義，因此在不同人眼裡一個考試的公平性與否的結論可能是大相逕庭的。一個考試技術意

義上的公平首先是同等對待所有考生，例如，考場或全部有空調或全部沒有，在後者的情況下有空調的考場也不許開。這種公平更主要地體現在考試內容的設計上。技術公平的準確含義是，如果兩個考生（組）在考查的心理特質上的水準相同，則他們的考試分數就應相同，而不受他們屬於哪個特殊的考生組別的影響。客觀地說，這種技術上的公平性尚容易把握，這也是該標準花了很大筆墨強調、從業人員應該花大力氣做到的。較複雜的是社會意義上的公平，如對不同民族人種來說考試通過率是否應該相等或近似，如果低分數是由於學生缺乏學習機會和條件造成的是否應該認為是不公平的，這幾乎觸及了美國政治的敏感神經。某種意義上說，該標準對這類敏感問題除努力做到「政治正確」外，並未提出完美答案。

在分數使用上，該標準特別注意考生的權力和責任。要求：所有有關考試內容和結果使用方法的資訊都應該對考生「透明化」，涉及考生個人資訊的資料使用要符合法律規定，考試機構必須保護考試資料，考生應該在考試之前被告知舞弊的後果和處理方式。殘疾考生的權利應受到特別的保護，考試機構有義務為殘疾考生提供考試方式的調整。基於美國「民族熔爐」的事實，該標準甚至對不同語言背景的考生權利也做了規定。

貳、先進考試機構或考試專案在考試分數產生、報告、解釋和使用上的技術現狀

一、分數的產生

在著名的考試評價專案中，很少能夠見到直接報告和使用考試的原始分的，而是大都要對原始分進行一系列線性或非線性的轉換獲得所謂「量表分數」（Scaled score）或「匯出分數」（Derived score）。同一考試不同試卷版本間的「等值」或「可比性」幾乎是任一社會化考試的必要條件。如美國大學聯合

會（http://www.collegeboard.org）擁有、ETS（http://www.ets.org）承辦的 SAT 考試，考生在每個試卷版本上得到的原始分數稱「公式分數」（Formula score），考生答對一題得一分，答錯一題則要減去四分之一或五分之一分（公式分數因此得名）。但這並不是最後告知考生的分數，由於這個公式分數受不同試卷版本的難度影響，參加不同年度或月份考試的考生並不可以直接比較，因此，它要被通過複雜的設計和統計運算轉換到一個最小值為 200、最大值為 800、單位為 10 的量表上去，這個量表將 1995 年考生群體分數的平均值規定為 500，標準差規定為 100，同時分數的使用者如大學的招生相關人員還可以從考試機構提供的一個對照表中查出每個分數所對應著的相應百分等級（http://professionals.collegeboard.com/data-reports-research/sat/equivalence-tables）。

二、分數的報告

分數的報告方式很大程度上取決於技術條件的許可，但也反應出教育思想和考試理念的變化。如果粗略地劃分一下，大概可以分作下面幾代：第一代是傳統的用信件或佈告通知考試分數的傳統做法。第二代則是由於技術手段的進步進行電子化報告，手段和技術的進步與觀念和思想的進步之間的關係有時候很難分得清誰在影響誰，在第二代分數報告中，一個熱門話題是所謂「診斷性」，考試機構提供的資訊不僅限於幾個考試學科的得分，而是深入到了「次級分數」（Sub-score），考生所獲得的不光是一個總分或「通過/不通過」的簡單結論，而是伴隨著如何得到這個結論的一些深層原因的推測和判斷，診斷性資訊使考試結果的作用得到了極大的擴展，本來設計用做選拔的很多考試開始兼顧到了幫助考生發展和幫助改進教學。技術和理念互相促進，考試形式開始出現了很多讓人意想不到的變化。目前的主流報告方式可以算作第三代，以美國大學聯合會擁有、ETS 承辦的另一個考試 AP 為例（https://scores.collegeboard.org/pawra/home.action）。全世界參加 AP 考試的每

一個學校被授予一個學校代碼，每個校長獲得一個校長代碼，校長可以為轄下的老師分配次一級代碼。每個老師可以在網路上看到他/她所教的每一個學生在考試中的具體表現，他/她可以把這種表現和全世界的學生做對比，去尋找和定位自己教學中存在的問題。歷次考試的資料被以時間為軸管理起來，形成目前 IT 行業中所稱的「資料倉庫」，反映了考試成績或教育教學的結果隨著時間的變化趨勢。第四代報告方式目前正在走出實驗室，它以資料倉庫和資料採擷技術為基礎，將考試和學校的日常教學和管理系統融為一體，起到類似於「汽車儀錶盤」的作用，隨時診斷、改進日常教學，及時肯定正面因素，發現和警報負面因素。

三、分數的解釋

如果把 Glaser（1963）列為考試行業的經典文獻之一恐怕提出異議的不多。Glaser（1963）第一次提出了「常模參照性考試」和「標準參照性考試」的概念，它幾乎成為了本領域最基本和最具普及性的概念之一。上世紀八十年代的高考標準化改革幾乎就是從普及「高考是標準參照性考試」開始的。然而，隨著技術和觀念的變化，人們現在更傾向于把常模參照性考試和標準參照性考試看作一個連續統（準確地說，連續統是一個嚴格的數學概念，此處取其連綿不斷之意）的兩端，很少有考試是絕對的常模參照性考試或標準參照性考試，大多數情況下，他們是兩者的混合體。更有甚者，持絕對觀點的人認為所謂「常模參照」和「標準參照」應該是指兩種分數解釋方式（Frisbie, 2005），他們不是兩種不同性質的考試。基於一個考試，可以既有常模參照的分數解釋，也有標準參照的分數解釋，而且，在很多情況下，這是完全必要的，只有這樣做才能更充分地發揮考試的功能。以 ETS 的 TOEFL 考試為例，幾乎所有人都知道其設計目的是為說明非美國學生到北美上大學或讀研究生用的，其手冊上也明文標明該考試是常模參照性考試。但是，2010 年，ETS 宣佈 TOEFL 考試將增

加報告「藍思分數」（http://www.ets.org/toefl/ibt/prepare/lexile/）。藍思分數絕不能算作常模參照分數。美國兩個類似於中國高考的考試 SAT 和 ACT 近些年也分別開始嘗試建立所謂「基準分數」（Benchmark scores）。這個分數的作用是告訴大學招生官員和考生及其家長該學生是否已經做好了上大學的準備（在大學基礎課學習中獲得 C 以上或 GPA 達到 2.76 的概率），這和「常模」也已經幾乎沒有關係，而越看越像一種「標準」了。

四、分數的使用

談到考試分數的使用，不能不提到高考分數使用中兩種普遍存在的現象：誤用和濫用。如某省教育廳網站上聲稱「本省一本上線率較去年提升 5%，說明我省基礎教育品質有很大提高」，升學率提高 5%的原因其實是大學擴招，與教育品質何干？由於採用原始分，不可能準確控制試卷難度，高考分數年度間難以避免地存在「深一腳、淺一腳」的現象，把「平均分」、「上線率」拿來評價教學品質，其作用簡直是誤導。因此，政府公函不止一次明文規定學校或地方政府「不許用考試成績排隊」，某省甚至通過政府檔強硬規定「不許公佈高考成績」，被媒體諷為「把高考成績捂成了國家機密」（http://www.cnjxol.com/epaper/jxrb/html/2011-06/24/content_479352.htm）。但是，用成績排隊其實是考試的自然結果，如果正確地理解「不許排隊」所表達出的對誤用考試結果，不顧學習背景和條件只看最終成績的不科學做法的批評，考試機構和研究者卻沒有及時告訴學校或地方政府除排隊之外還有什麼做法能科學使用考試成績評價教學績效，這其實是非常合理的要求。本節本應介紹國外的先進經驗，卻說起了我們自己的問題。確實，在分數使用上，我們面臨的很多問題是非常中國化的。

與西方國家相比，無論是考試機構，還是各級研究者，嚴重缺少對高考效度提供有力證據的研究。同樣以 SAT 為例，2006 年大學

聯合會收集了多達 110 所大學共 151316 名學生的一年級 GPA 資料，分析 SAT 對大學一年級 GPA 的預測能力，所付努力可見一斑（http://professionals.collegeboard.com/data-reports-research/sat/validity-studies）。

參、「雲海工程」的設計藍圖

「雲海工程」指大陸教育部考試中心在雲南、海南兩省對全國普通高等學校入學統一考試（高考）在分數產生、報告、闡釋、使用等技術環節進行的一籃子試點。試點工作從 2010 年起，首先在海南省進行，2011 年擴大到雲南省。

試點之初，適逢《國家中長期教育改革與發展規劃綱要》頒發，「綱要」為試點提供了最強有力的依據和支援。綱要要求「完善專業考試機構功能，提高服務能力和水準」。這是國家級的教育改革指導檔中第一次明文規定考試機構的任務。尤其是「專業」二字，明確了考試機構不同於通常觀念中政府部門的定位。據利伯曼（M. Lieberman）給出的定義，專業（profession）應具備以下基本特徵：一、範圍明確，壟斷地從事社會不可缺少的工作；二、運用高度的理智性技術；三、需要長期的專業教育；四、從事者無論個人、集體均具有廣泛的自律性；五、專業的自律性範圍內，直接負有作出判斷、採取行為的責任；六、非營利，以服務為動機；七、形成了綜合性的自治組織；八、擁有應用方式具體化了的倫理綱領。以此觀之，我們距離「專業」甚遠。

「雲海工程」是考試機構轉變觀念，將考試定位為公共服務，將考試的各種服務物件如考生、大學、中學等視作消費者，從而很自然地對傳統考試重選拔輕評價、重結果輕過程、重管理輕服務、重政策輕資料的狀況作出反思和改變。通過大量的理論研究和調研，「雲海工程」把高考在分數產生、報告、闡釋和使用方面的任務和目標確定為：一、重新設計和優化考試報名環節，收集考生的社會、經濟、文化、教育等背景資訊，探索利用這些資訊和考試結果進

行綜合評價。考試結果的闡釋必須結合考生的背景資訊進行才能做到更加科學、合理、有效。報名是收集考生背景資訊的最好時機。因此，「雲海工程」首先從考生報名環節入手，將高考錄取必須的考生資訊定義為核心報名資訊，將對於教育管理和決策有益的資訊定義為擴展報名資訊。核心報名資訊為考生必填，擴展報名資訊為考生選填。同時，還專門設計了考生問卷，在社會經濟背景、學習方式條件和社會關心的熱點問題等方面對考生進行問卷調查。二、探索對原始分進行標定、校準等轉換，逐步形成科學、準確、易於解釋和使用的量表分數。上世紀八九十年代高考曾探索過「標準分轉換」，高峰期曾有多達三分之一的省區使用標準分錄取。後來，由於強大的傳統觀念和本身的技術缺陷逐漸下馬，至今只有海南頑強堅持。現在，雖然很多業內人士承認和肯定「標準分」的價值和意義，但這一教訓和很多類似的歷史故事都提醒我們面臨傳統時科學理論和實踐進步的艱難。因此，「雲海工程」在設計時避開了容易捲入政治和利益問題的爭論，確定了「暫時不和錄取掛鉤」的基本思想。從研究入手，從利用考試結果改進教學和教育管理入手，不影響考生切身利益，換取自身存在的權利，寄希望於逐漸影響人們的思想和習慣，寄希望于未來。應該說，這種柔性姿態，和長期以來習慣於政府主導一切的傳統相比，是一個巨大的進步，體現了對自身定位的轉變。三、探索實現不同次考試間考試分數的等值，作為開展評價工作的基礎。高考號稱國家統一考試，其實從 2003 年之後，已經有半壁江山實行「分省命題」，嚴重缺乏「分數的可比性」。這種可比性一是對相同省區市來說在不同年度間缺乏可比性，這直接造成了長期追蹤教育品質發展變化趨勢的困難，每年都重新歸零，不能充分發揮高考在教育績效評估上的作用；一是在採用不同試卷的省區市之間缺乏可比性，造成國家對地區差別缺乏瞭解，招生指標和計畫在地區間調整困難。因此，同一地區年度之間的「縱向等值」和不同地區同一年度的「橫向等值」都是必須解決的問題。四、探索提供診斷性分數報告。考試分數的診斷功能是發揮大規模考試對教學

正面導向作用的技術前提。網路技術的發達和普及使考試資料的深入挖掘和資訊交換成為可能。要改變傳統分數報告「一張紙條上幾個分數」的做法，正面提供使用考試結果改進教學的範例，引導中學和教育行政部門正確、合理地利用考試結果。五、探索提供標準參照性分數解釋。所謂標準參照，就是在進行分數解釋時，不是著眼於考生之間的位置，而是和某一事先確定的外部標準去做比較。例如，通過一個專家組的工作，為一張試卷確定一條或幾條具有外在意義或功能的關鍵分數線，判斷考生是否達到了這些關鍵分數線。對於達到了某些分數線的考生，以往考試機構通常還習慣用一些描述性的語言來說明考生具體的水準，如「有較好的分析問題的能力，能比較熟練地運用所學知識，但有一些環節處理還欠妥當」等等。這些描述性語言的主要功能是從「一般」的角度回答獲得某個考試分數的考生「能做什麼」。但是，這種描述性語言大多語焉不詳，包含大量缺少定量資訊的形容詞，因此在教學活動中教師和學生很難根據這些描述採取恰當的對策。為此，一種替代策略是從「特殊」的角度看問題，拿一些典型的題目作為例子，給出特定水準的考生在這些題目上的答對概率或期望得分，從而幫助教師和學生更加具體地去理解考試分數的涵義。這種「特殊」和「一般」的策略可以交替使用，一般來說，有代表性的題目和特定水準的考生在這些題目上的答對概率是描述性的「能做什麼」的具體化，而「能做什麼」則是對題目和答對概率的推廣，兩者可以互相配合。在初始階段，以題目和答對概率為主，在積累了足夠的資訊後，則可以比較準確詳實地給出「能做什麼」。六、探索「增值評價」的技術和方法，跳出單獨一次考試的局限，嘗試建立跨越時間和空間的連續的考試評價體系，打破「一次考試定終身」。七、在以上這些技術準備的前提下，建立基於資料採擷技術的資料倉庫，向考生、中學、大學、教育行政部門乃至命題部門提供個性化評價報告。資料倉庫是起源自 IT 行業的一個熱門名詞。它和目前我們在業務工作中廣泛使用的資料庫不同。資料庫一般是針對某個技術流程和環節的，如報名資料庫是關

於考生的報名資訊的，考試資料庫是關於考生的答題情況的，而資料倉庫一般是以時間為主題的。以往我們招生考試工作的習慣是一年一結，當年的資料對第二年毫無影響，因此當年工作結束後就全部清空。資料倉庫則是將每個年度的資料按時間為軸組織起來，著眼於發現趨勢和規律性的東西。考試機構的責任是將紛繁雜亂的資訊加以整理規範，提煉出有價值的規律和趨勢，然後以通俗易懂的方式報告給正確的物件。

同時，「雲海工程」是調整國家考試機構和省級考試機構的關係、在新的形勢下建立國家考試機構和省級考試機構合理工作模式的嘗試。我國的國情與美國等西方國家不同，不存在像 ETS 那樣的考試公司。我國考試的基本管理模型是中央和地方分級管理。在傳統模式下，中央級的機構主要負責政策，省和省以下的機構主要負責實施。從建立專業考試機構的要求來看，這種模式存在很大的弊病。和傳統經濟中的實物流或金錢流一樣，現代經濟中的資料流程是業務運行的血脈。傳統模型容易造成資料孤島，資料堵塞在一些節點難以流動，資料採擷無從談起。「雲海工程」中摸索的考試中心和省級考試機構之間新的工作模式的大概雛形是：一、考試中心和省級考試機構共同確定各個年度的目標和任務；共同制定各個工作環節的資料標準。二、省級考試機構根據標準採集和準備原始資料；三、考試中心統一負責理論和技術的研發，提出技術問題的解決方案和主要演算法，開發應用軟體原型；四、省級考試機構上報原始資料，由考試中心統一完成一部分需要使用 IRT 理論等比較複雜的測量技術模型的資料計算，考試中心向省級考試機構返回中間階段的輸出結果；五、省級考試機構根據考試中心提供的中間階段輸出結果完成後續計算；六、省級考試機構負責面向考生、中學、大學等直接用戶發佈和解釋結果，考試中心負責向命題部門發佈和解釋結果。

肆、艱苦的改革探索

實事求是地說，目前高考改革的環境非常惡劣。一方面，領導和群眾都對高考的現狀非常不滿，強烈要求改變；另一方面，由於高考是多種社會和教育矛盾的聚焦和妥協，任何一點改變往往迎來的是批評大於讚揚。考試中心兩任領導班子支持和鼓勵改革，表現了他們對宏觀形勢高超的把握能力和對考試技術發展方向的準確判斷。海南省從 2010 年起步開始試點，雲南省從 2011 年起步開始試點，表現了這些地方考試機構的領導者不甘平庸、敢於創新的勇氣和負責任的精神。

在試點過程中，考試中心和相關省考試機構首先建立了高考資料分析的基本規範和標準程式。高考資料分析和處理分兩個階段，分別稱作「初始分析」和「最終分析」。「初始分析」完成於高考成績正式公佈之前，它有兩個主要作用：一是及時發現題目答案設定和評分過程中可能的錯誤；一是利用部分分析結果產生面向考生和大學的分數報告內容。由於要直接面向考生，這是整個工作過程中時間要求最緊、準確程度要求最高、風險最大的環節。尤其對考試中心來說，從省考試機構完成評卷到成績發佈，只有三到四天時間，考試成績必須準時發佈，否則就會引起社會的不安。「最終分析」在高考成績公佈之後至秋季學期開學前時間比較寬裕的條件下進行，它的主要作用是產生各種後續評價報告，為命題和考試決策人員、中學教學人員、教育行政管理人員提供參考資訊。

初始分析主要內容包括：基本描述統計量（試卷平均分、試卷標準差、分數分佈長條圖等），題目的經典難度和經典區分度分析，選擇題的選項功能分析，各個科目之間的相關係數矩陣和每個科目內各個考查目標間的相關係數矩陣，IRT 基本分析，總分和各個科目的原始分與百分等級分數的對應關係，每

個考生在各個題目和各個次級考查目標上的實際得分率和期望得分率。最終分析首先要獨立重複一遍初始分析的內容，還要包括：分類描述統計量（將考生按各種自然和社會資訊分成不同的類別，對各個類別計算描述統計量）、題目功能差異（DIF）分析、各個科目的信度和測量標準誤分析、探索性因素分析和驗證性因素分析、聚類分析、多維尺度分析、等值分析。

從技術上看，雖然以上涉及的大多數技術環節在發達國家的理論研究和實踐應用中都有很多成功範例，但中國確實有自己獨有的國情。中國家庭和學校對待考試的態度是西方人無法比擬甚至不可想像的，很多在美國等國家行之有效的辦法在中國行不通。如等值資料收集設計，在 SAT、NAEP、TOEFL、PISA 等考試評價專案中通行的做法在我國均會碰到問題。曾經有過令人沮喪的例子，考試中心兩個工作人員帶著試題到一個偏遠山區進行等值試測，為減少當地學校對試題的關注，事先告訴他們工作人員是省考試院的，學校校長微笑點頭，告別時卻說「我知道你們不是省考試院的，這些題目多半將來是要派上用場的，我們都複印下來了。」工作人員好奇之下問他是如何判斷的，他笑著說「你們都說的是普通話。」

幾個主要的技術環節中，等值的主要挑戰來自資料收集設計，目前有兩種資料收集設計正在試驗。一是在每年 5 月初（距高考還有一個月）從每個省選取 600 名左右水準分佈均勻的考生進行錨題測試；一是在每年大學秋季開學後利用大學摸底考試的時機進行錨題測試。兩種資料收集設計較國外常用做法均有創新。診斷性分數報告的挑戰則主要來自這種報告技術本身。診斷性分數報告是目前世界上研究和應用熱點，國外很多著名考試專案都在力求有所突破。和他們相比，我們面臨更多的困難，尤其是考查內容過多，題目數量太少。如理科綜合，內容橫跨物理、化學、生物三個中學學科，如何合理確定子分數的維度以保證足夠的信度，不光需要大量的積累和探索，最終更需要命題和評價的互動。IRT 的應用也碰到一些問題。國外考試多為選擇題等客觀性試題，即

使有寫作等主觀題，涉及的分數等級也很少，如習慣上美國的作文多採用 1 到 6 的分數等級評分，而高考作文的分數等級赫然達 61 個之多（0 到 60），以致商業化 IRT 軟體無法處理。這種計分辦法更多地出於習慣，但習慣往往是難以改變的。增值評價的理論和技術手段非常豐富和成熟，但增值評價的實施需要牽涉到不同主管部門，在「條塊分割」的管理體制下真正推廣應用需要體制和機制創新。

「雲海工程」確定將「普通高考評價資訊回饋和服務網站」作為試驗結果的集中展示平臺。按照不同的服務物件，網站可以細分為以下主要內容。一、面向考生個人的成績報告和分析網站。成績報告指把考生關注的得分情況通過網路準確、及時、便捷地通知考生及其家庭。成績分析則是對成績資料進行更深入的分析和整理，告知考生的長處和缺陷，幫助考生從更多維度認識自己，選擇專業和發展方向。二、面向中學和教育行政部門的資訊服務。幫助他們科學、合理、有效地利用高考結果中蘊藏的資訊改進教育教學，改進教育管理。三、面向大學的服務目標是從考試技術方面為大學將來根據「規劃綱要」的要求實行「多元評價、綜合錄取」創造條件，多角度多層面提供考生的資訊。同時也幫助大學利用歷年招生和考試結果中的資訊研究存在的問題，改進招生辦法。在現行管理體制之下，考試中心和大學基本上不發生關係。通過主動為大學提供服務，考試中心爭取未來能建立採集考生進入大學後學習表現資料的管道，以進行高考效度分析，為改進高考服務。

面向考生的網站以省級考試機構為責任主體，考試中心提供技術支持和幫助。面向中學和教育行政部門的網站因為涉及一系列理念和技術問題，省考試機構缺乏專門人才和研究，故由考試中心統一設計和開發，省級考試機構管理和使用。該部分內容由於突破了傳統觀念高考用於評價高中教學的禁區，為避免引起爭議，在技術上進行了充分的考慮，初期只提供容易獲得共識的服務，甚至不提供每所中學的「平均分」等常見資料，以避免引起基礎教育部門誤解。

由於高考改革整體方案還不明朗，面向大學的網站除把考生個人資訊中新增加的內容提供給高校之外，無論從設計和技術上都還不成熟，故而 2011 年暫未進行實際試點，等待相關政策的進一步明朗。

伍、未來與方向

2011 年高考結束後，中央電視臺等新聞媒體對「雲海工程」試點情況組織了專題採訪，對試點工作做了充分肯定。

2012 年 4 月，應美國教育測量學會 NCME 主席琳達·庫克女士邀請，考試中心在加拿大溫哥華舉行的 2012 年 AERA 和 NCME 聯合學術年會上舉辦了特邀專題「考試在中國：分數報告、闡釋和使用的改革探索」。中國考試工作者的改革努力受到了世界同行的高度評價和認可。

2012 年初，考試中心對試點工作進行了總結，實事求是地分析了試點工作情況，在充分肯定成績的前提下，重點討論了存在的缺陷和不足。認識到儘管在事先反復強調「和錄取暫不掛鉤」，但無論是大學、省考試機構，還是考生，或許出於對改革目標的認可，或許出於功利心態，存在急於和錄取掛鉤的傾向。特別是嚴重缺乏資訊管道的農村偏遠地區考生，非常容易受到專業性向測驗等結果的影響，缺乏科學使用心理測驗結果的條件，容易產生片面性和副作用，造成評價結果影響考生個人利益等欲速則不達的局面。為此，考試中心確立了「積極地發揮考試結果在指導中學教學上的作用，謹慎地發揮考試結果在評價教育品質上的作用，持續地研究考試結果在促進考生個體發展上的作用，同時限制在高利害決策中的實際應用」的基本原則，在 2012 年試點工作中，考試中心相應地作出了政策和技術調整，提出集中力量，以群體評價和自身評價為主要政策突破口，以等值為主要技術突破口，儘快實現分省命題和全國命題考試分數的年度間和地區間的可比性，提供政府宏觀決策參考依據。首先著

眼於考試自身，解決好高考長期以來存在的一些技術缺陷，同時更加強調評價和命題的一體化，形成「評價為命題服務，命題促進評價」良性互動局面。

目前，我國大多數省陸續進入新課程改革後的高考，又帶來了一些新的問題，如作為新課改後高考重要特徵之一的「選考模組」要求考生從若干個備選題目中選做一部分，這給傳統的統計和分析方法帶來了新的挑戰。

「規劃綱要」將高考內容和形式改革列為高考改革在技術層面的主要任務，要求「著重考查綜合素質和能力」。技術是手段，不是目的。高考改革的最終目的是「以考試招生制度改革為突破口，克服一考定終身的弊端，推進素質教育實施和創新人才培養」（見「綱要」）。作為專業考試機構，要抓住高考內容和形式改革這個重點和難點，緊緊圍繞這個重點和難點，推進技術創新，為體制機制創新提供條件。

參考文獻

American Educational Research Association, American Psychological Association, National Council on Measurement in Education [AERA/APA/NCME]. (1999). *Standards for educational and psychological testing*. Washington, DC: American Psychological Association.

Frisbie, D. A. (2005). Measurement 101: Some fundamentals revisited. *Educational Measurement: Issues and Practice*, *24*(3), 21-28.

Glaser, R. (1963). Instructional technology and the measurement of learning outcomes: Some questions. *American Psychologist*, *18*, 519-521.

從寫作題型與閱卷機制探討臺灣高中畢業生英文寫作能力的評量

張武昌 [1] 林秀慧 [2]

中國文化大學 [1] 大學入學考試中心 [2]

摘要

本篇論文包含兩個部分，首先簡介台灣歷年大學入學考試出現過之重要寫作題型，並探討這些題型的測驗目標及在教學面產生的回沖效應。大考中心對於英文寫作能力測驗的研發一向不遺餘力，也針對不同程度考生研發了包含句子層次、篇章層次與段落層次等多種寫作題型，以評量其寫作能力。在此部分也針對大考中心所推出的「英文寫作能力測驗」及所涵蓋的題型與測驗目標提供了簡略的說明。

論文第二部分則聚焦於考生試卷評閱的相關議題，包括評分標準、大考閱卷的標準流程（評分標準訂定、試閱會議、正式評閱），以及閱卷者訓練等。大考中心自 2001 年開始研發電腦螢幕閱卷系統，同年英文考科於「英文寫作能力測驗」中首次採用此系統進行閱卷，並於 2002 年與 2003 年兩次學測補考以改進後之系統閱卷。2008 年配合英文寫作能力測驗研究計畫案進行分題電腦螢幕閱卷，並且針對高中教師參與閱卷工作的可行性進行初步的研究，以上議題在本論文第二部分皆有簡要的討論，並針對未來進行相關研究時應該注意的事項提出建議。

關鍵詞：英文寫作能力測驗、寫作題型、閱卷、電腦螢幕閱卷

張武昌，中國文化大學外國語文學院院長
林秀慧，大學入學考試中心高級專員

Assessing Writing Proficiency of Taiwanese High School Graduates: Tasks and Marking Schemes

Vincent W. Chang[1], Hsiu-Hui Lin[2]

Chinese Culture University[1], College Entrance Examination Center[2]

Abstract

This paper consists of two parts. In the first part, major tasks that have been employed over the years to assess senior high school graduates' writing proficiency in Taiwan's college entrance exams are presented and discussed. These tasks are essentially of two types: (i) free composition and (ii) guided composition, which includes tasks involving the situational-writing format, two-topic-sentence format, picture-prompt format, etc. Writing tasks designed for College Entrance Examination Center English Writing Proficiency Test (CEWPT) are also presented in the first part. The second part of this paper focuses on the marking of examinee scripts, including the rating scales, the marking procedures, and the training of raters. On-screen marking of examinee scripts was first conducted by the College Entrance Examination Center in 2001, and after this trial run, the new marking method has thus far been adopted five times for marking examinee scripts, twice in the make-up exams of the General Scholastic Ability Test (GSAT). Raters' viewpoints regarding on-screen marking and suggestions for its future use are also presented in the second part of this paper.

Keywords: College Entrance Examination Center English Writing Proficiency Test (CEWPT), English Writing Proficiency Test, Marking/Rating Scheme, Writing Types, On-Screen Marking, Writing Tasks

---

Vincent W. Chang, Professor and Dean, College of Foreign Languages Chinese Culture University
Hsiu-Hui Lin, Senior Staff Member, College Entrance Examination Center

I. Writing tasks in Taiwan's college entrance exams and the CEWPT

Assessment of high school graduates' writing proficiency in Taiwan's college entrance exams began in 1974, with multiple-choice items as the assessment technique (Huang 1994), and it was not until 1981 that the examinees were given a direct writing task requiring them to produce an English composition of 60 to 80 words, beginning with the topic sentence: *I have long hoped to become a college student*. The examinees were also required to write the composition based on a set of vocabulary items provided, and penalty was imposed if they failed to use the vocabulary items:

Starting with the topic sentence provided, write an English composition of 60 to 80 words, exclusive of the nine words in the topic sentence. Use at least 10 of the following 12 vocabulary items in any order and tense, and underline them in your composition. Two points will be deducted if you use only nine of them, four points if only eight of them, and so on. The total score is 20 points.

Topic sentence: I have long hoped to become a college student.

Vocabulary items:	1. make up my mind	2. as early as	
	3. specialize	4. on the one hand	
	5. on the other	6. enable	
	7. after all	8. moreover	
	9. do my best	10. lose heart	
	11. certainly	12. success	(JCEE, 1981)

It is apparent that the examinees' performance must have been severely

hindered by the imposed restriction on vocabulary use. Furthermore, the marking of the examinees' compositions must also have been a very tiring job, since the raters had to constantly check back and forth whether or not the required vocabulary items appeared in the compositions (Chang, 2000). It is perhaps because of these reasons that later writing tasks in the college entrance exams abandoned this highly restrictive format and in the three years that followed, free composition was adopted as an assessment measure of examinees' writing ability.

From 1982 to 1984, the examinees were required to write a composition of approximately 80 words on the topics of "The National Flag and I," "A Taxi Ride," and "How I Spent Yesterday Evening," respectively. However, the topics of free composition should relate closely to the experience of the students and as a general rule, abstract topics should be avoided (Heaton 1990). It is apparent, from this perspective, that although the three composition topics relate in different ways to the examinees' daily experience, they fall short of qualifying as realistic writing tasks, especially the first one. "The National Flag and I" is simply too abstract a topic for high school graduates to write anything meaningful and genuinely communicative on. Another weakness with free composition, as Huang (1994) pointed out, is that quite a number of examinees were found to have committed to memory many model compositions on a variety of topics when they were preparing themselves for the English test of college entrance exams. This explains why since 1985 writing tasks in the English test of Taiwan's college entrance exams have abandoned free composition in favor of guided or situational compositions.

The writing task of the 1985 English test, for example, required examinees to write a composition on the topic of "Leisure Activities," with the explicit instruction

below to guide the examinees in their writing:

> Write briefly in English about a leisure activity (for example, traveling, physical exercise or sport, going to the movies, listening to the music, or reading books you love, etc.) which you always wanted to do during high school days but for some reason(s) couldn't and now hope to do right after being admitted to college. Divide your article into two paragraphs, the first explaining why you couldn't do it and the second stating how you plan to do it. The length of your composition should be about 80 words. (Translation from the Chinese original)

It is worth noting that this writing task required the examinees to divide the composition into two paragraphs, each focusing on a different aspect of the topic. With this explicit instruction, a writing task involving the format of "Situational Writing" reduces to a large extent the possibility that the examinees may simply memorize a number of model compositions for use in the entrance exams. However, it was soon found out that the explicit instruction provided more often than not led the examinees to turn the writing task into translation. That is, the instruction given in Chinese, originally intended to serve as a general guideline, was nevertheless translated, almost verbatim, into English by a great number of examinees, and the translation usually constituted a significant portion of each examinee's composition. It is mainly for this reason that starting in 1992 writing tasks in college entrance exams adopted a new format: Composition with Two Topic Sentences, as shown below:

> Write an English composition of about 100 words on the topic "Time." Divide your composition into two paragraphs; start the first with the topic sentence "Lost time

is never found again" and the second with the topic sentence "Now I have a new plan for using my time wisely."

Examinees of the 1992 English test were required to write about 100 words—20 words more than was required in previous years—on the topic of "Time" and in the two paragraphs they were required to write, a topic sentence was explicitly provided for each paragraph. Each topic sentence served as a guidepost for the examinees to generate ideas related to the assigned topic and to organize the ideas in a logical and coherent way. This two-topic-sentence format avoided the problem with explicit instruction in situational writing, because there was nothing in this writing task for the examinees to translate into Chinese.

Since 1994, high school graduates in Taiwan have been given an additional channel through which they can be admitted to colleges and universities. This is commonly referred to as the General Scholastic Ability Test (GSAT). Over the years, writing tasks in the English Test of the GSAT (GSAET) have been constructed in the format of guided composition, and examinees have been given a variety of topics to write on. The trend has been to assign realistic, communicative writing tasks, with proper prompts, for examinees to express their views on the given topics. Take, for example, the writing task of the 1995 GSAET. The task required the examinees to write, on behalf of a Chinese high school student, Chih-ping Wang, an English letter of about 100-150 words to his American pen-pal George, who planned to come with his parents to Taiwan to stay for two years or so. An explicit instruction was given regarding what to include in the letter, as shown below:

In your letter, you should first extend your welcome and then offer your

suggestions regarding the question George raised in his letter: "Can you give me some suggestions regarding what I should do and what I should not do when I am in Taiwan?" (Translation from the Chinese original)

This writing task followed closely the requirement specified in the Curriculum Guideline that high school graduates should be able to write simple letters to inform or request information, and thus came close to being genuinely communicative. Writing tasks in the GSAET typically require the examinees to produce writing that is closely related to daily life, and they are provided with the needed prompts, verbal or non-verbal, to complete the task. The prompt for the writing task of the 1998 GSAET, for instance, was a short poem written by Langston Hughes:

April Rain Song

Let the rain kiss you.
Let the rain beat upon your head with silver liquid drops.
Let the rain sing you a lullaby.

The rain makes still pools on the sidewalk.
The rain makes running pools in the gutter.
The rain plays a little sleep-song on our roof at night—

And I love the rain.

The examinees were required to first answer five questions based on their understanding of the poem and then write an English composition of about 120 words. The questions and the rubric for the writing task are as follows:

Comprehension questions:

1. Which season of the year serves as the setting of the poem?
2. Which word in the poem is closest in meaning to sleep-song?
3. What does the phrase silver liquid drops refer to?
4. Which word in the poem is opposite in meaning to running?
5. Which of the following words best describes the rain in this poem: *boring*, *harsh*, *depressing*, *heavy*, *hopeful*, or *musical*?

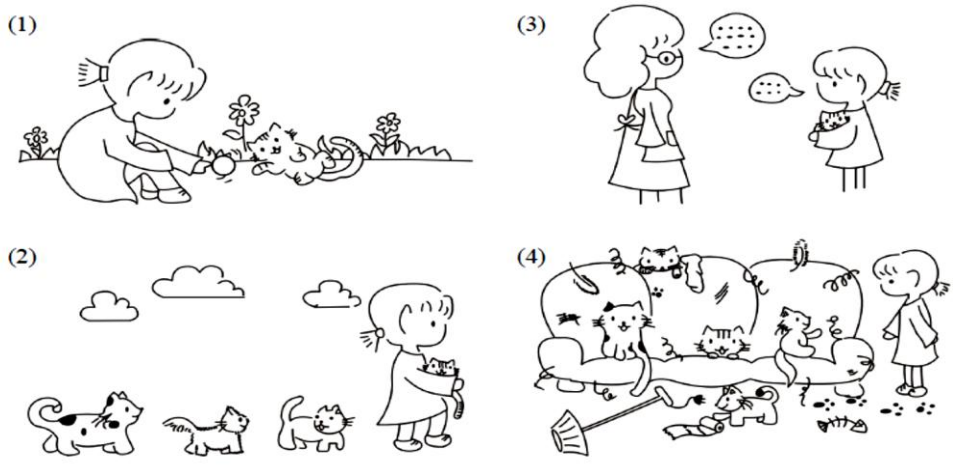
The rubric:

> Everyone feels differently about the rain under different circumstances. First describe an event you actually experienced or a scene you witnessed on a rainy day. Then write to express how you feel about the rain based on the event or the scene. (Translation from the Chinese original)

This task received positive feedback from both teachers and students, not only because the examinees were presented with a meaningful, realistic writing task but also because it is in essence an "integrative" test measuring the examinees' reading as well as writing abilities.

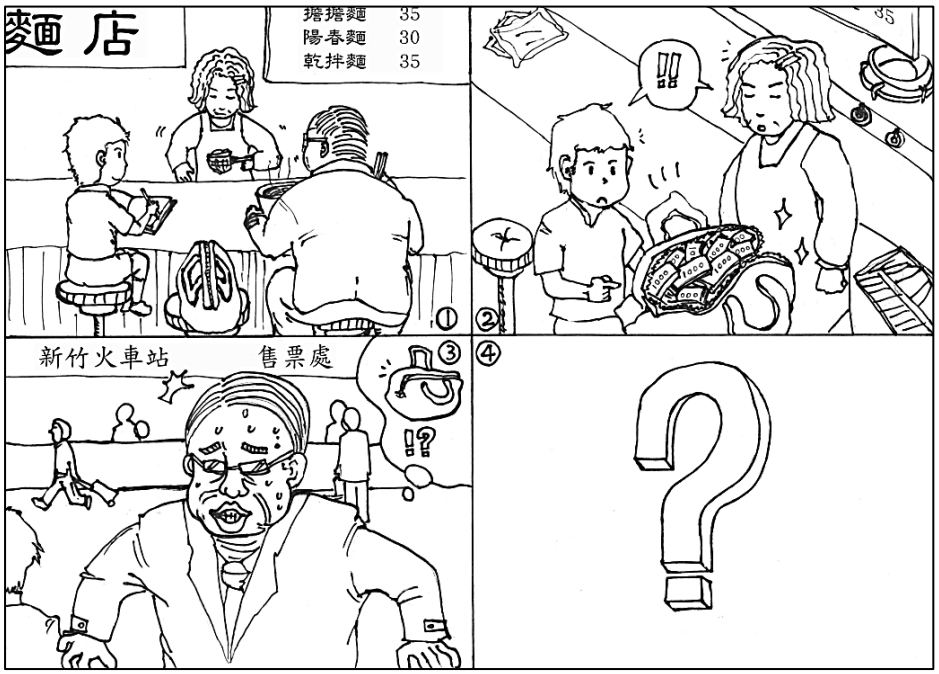
Non-verbal prompts for writing tasks in Taiwan's college entrance exams first appeared in the 2004 GSAET. The examinees were required to write an English composition of approximately 120 words, beginning with "One evening…" In this task, the examinees were shown three pictures, with the first showing a person attending his friend's wedding banquet (and perhaps drinking one glass too many). The second picture shows the drunken man trying to hire a taxi home, and the third picture shows the man awoke to find himself in a police station.

Ever since its first appearance, the picture-writing task has received warm welcome from both teachers and students not only because it provides vivid pictorial prompts for the examinees to write the composition but also because it allows them to stretch their imagination and show their creativity when describing what they think happens in the pictures. The writing task of the 2007 GSAET, adopting a four-picture format, went a step further and allowed the examinees to either follow the steps specified or rearrange the four pictures in any order as long as their story showed logical and coherent development.



The writing task of the 2010 GSAET, again adopting a 4-picture format, allowed even more room for the examinees to express their views concerning how

the story would develop to a logical ending. Here, the first picture shows a hard-working woman at a noodle stand, with a boy nearby writing his homework. The picture also shows a man eating noodles, with his handbag on a stool. The second picture shows the woman and the boy opening the bag—after the man left—and finding a lot of cash in it. The man is shown in the third picture panicking after arriving at the train station and finding that he has lost his bag. The fourth and last picture contains a big question mark, allowing the examinees ample room to come up with a coherent ending on their own.



Since 1998, the College Entrance Examination Center (CEEC) has conducted a series of research on the assessment of high school graduates' English writing proficiency. This research is of vital importance for two major reasons. At present the writing tasks in college entrance exams, including the GSAET (administered in February) and the DRET (Department Required English Test—administered in July), are the same for all examinees (Chinese-to-English translation and guided

writing). In a way, this ensures "fairness" among all examinees, but it apparently fails to take into account the fact that different departments may demand different levels of writing proficiency among in-coming freshmen, and there is not a specific level of writing proficiency required by all departments (Chang, 1999a). Examinees who wish to be admitted to the PE department of a certain university, for example, do not need to possess the same writing ability as those who wish to major in English or international trade. Tasks requiring the examinees to demonstrate memo- or note-writing ability are in fact more appropriate for applicants to the PE department than tasks, say, involving the reader-response type. It makes sense then to divide Taiwanese high school students' writing ability into different proficiency levels and design writing tasks appropriate for each level so that the students can take the test—whenever they are ready—to prove that they have achieved the proficiency level required for admission to the departments they hope to enter. As shown in Table 1, high school graduates' performance in the guided-writing task (total score=20 points) has not been satisfactory in both the GSAET and the DRET in recent years, causing a large number of high school graduates to feel enormously disappointed at their writing ability:

Table 1: Average scores on guided writing in college entrance exams (2007-2011)

year	**Guided Writing**	
	GSAET	DRET
2011	8.02	6.78
2010	7.11	6.87
2009	6.17	6.8
2008	7.08	6.42
2007	7.08	6.4

The CEEC English Writing Proficiency Test (CEWPT) can help solve the problem by allowing high school students to progressively challenge themselves when they believe they have achieved a particular writing proficiency level. Their success at a lower level helps to promote their interest in learning further and to challenge themselves at a more advanced level. The other, and more significant, reason for the CEWPT is that the CEEC has been under tremendous pressure because the marking of the GSAET and DRET examinee scripts must be completed within a very limited period of time (usually 7 to 10 days). If the CEWPT can be administered several times a year, and senior high school students are allowed to take the test any time they are ready, then the writing part can be removed from the GSAET and DRET, thus relieving the pressure of having to complete the marking within a limited time span. This also has the added advantage of increasing the reliability in marking since more raters can be recruited and trained, and better monitoring schemes can be implemented to ensure quality marking.

The CEWPT proposed in 1999 included two parts: (i) sentence-level writing tasks and (ii) discourse-level writing tasks. The former aimed at assessing examinees' knowledge of English sentence structures and their active manipulation of the structures. Four major writing tasks were included in this part: (a) sentence-combining, (b) sentence-making, (c) question and answer, and (d) detailed answers based on a reading passage. The discourse-level writing task, on the other hand, aimed to assess examinees' ability to write a coherent essay. This design was in accordance with the requirements stipulated in the Curricular Guideline that the teaching of writing skills in senior high schools should proceed from sentence-making, question and answer, paragraph-writing, letter-writing, to the writing of a coherent essay.

In sentence-combining, examinees' knowledge of different types of sentence structures (e.g., compound sentences, complex sentences, and compound-complex sentences) was assessed and the examinees were required to combine sentences with conjunctions of various types. The following example shows combining a pair of sentences with the subordinate conjunction *after*:

Sentence-combining:

a. Michael will go to college.

b. He graduates from high school. (*after*-clause)

→Michael will go to college after he graduates from high school.

The second task, sentence-making, aimed to assess whether or not examinees could accurately put some common, every-day words and phrases to effective use in writing. In the following example, the examinees were given the phrase *in fact* and were required to make a sentence containing this expression, as shown below:

Sentence-making:

Make a sentence containing the phrase *in fact*.

→Mr. Lin is not a professor; in fact, he is a salesperson.

In the third task, question and answer, examinees were given a question which they may encounter in their daily life and were required to provide a detailed (i.e., both structurally complete and sufficiently informative) answer to the question:

Question and answer:

Q: What do you want to be when you grow up?

A: I want to be a scientist because I have always wanted to know more about nature.

In the fourth task, examinees were assigned a reading passage, followed by a set of questions, and were told to provide detailed answers to the questions, for example:

Detailed answers based on a reading passage:

The passage:

The Tri Service General Hospital yesterday held a ceremony for 35 children who completed a hospital-sponsored weight-loss program during their summer vacation.

The students lost an average of 2.34 kilograms during the past two months, and the student who lost the most, the nine-year-old Howard Chang, *shed* more than six kilograms.

"My weight dropped to 48.3 from 54.4 kilograms, and my classmates won't be able to call me 'porker' anymore," Chang said happily.

"Howard calculated the calories of everything he ate during his participation in the program," his mother said.

"He would deliberate before eating even a slice of pizza because it has 350 calories," the mother said.

When asked what he most wanted to do following the accomplishment, Chang replied, "Eat at McDonald's."

Comprehension questions:

1. What was the ceremony for?
2. Why was Howard mentioned in the passage?
3. Why did Howard's classmates call him "porker"?
4. How much weight did Howard lose in the program?
5. How did Howard manage to lose so much weight?

The discourse-level writing task, in contrast, required the examinees to write a coherent essay of approximately 150-200 words based on a reading passage (for example, a fable). In the example below, examinees were required to (i) give a brief

summary of the fable, (ii) identify the moral of the fable, and (iii) write a separate paragraph detailing the inspiration they drew from the fable, with particular reference to their daily life experiences.

> A farmer, being on the point of death, called together his sons. Though he was not a wealthy man, he did want to show his sons the way to success in farming.
>
> "My children,"he said, "I am now departing from this life, but all that I have I leave you, and you will find it in the vineyard."
>
> The sons, supposing that their father was referring to some hidden treasure, set to work with their spades and ploughs and every implement that was at hand, and turned up the soil over and over again. They found no treasure, but the vines, strengthened and improved by this thorough digging, yielded a much finer vintage than ever before, and more than repaid the sons for all of their trouble.

II. Marking schemes for Taiwan's college entrance exams and the CEWPT

Prior to 1986, there was no explicit rating scale for the marking of examinees' compositions in the college entrance exams. The raters didn't have any specific guidelines to follow and seemed to "follow their heart" while marking the compositions. The first appearance of an explicit marking criterion was found in the rubrics of the 1986 writing task, as shown below:

Scoring criterion: Spelling and punctuation, 4 points; Diction, 4 points;
Grammar, 4 points; Organization, 4 points; Content, 4 points

This scoring criterion not only provided raters with a general guideline regarding what components of writing to attend to but also made explicit the weight of each component in marking examinees' compositions. This apparently had a

positive backwash in teaching as high school English teachers were made aware of what to pay attention to when teaching writing. More importantly, this criterion helped the raters to keep the scoring standards in mind while marking and thus helped to ensure greater inter-rater reliability. This criterion was adopted in the years that followed until its further modification in 1993:

Scoring criterion: Content, 5 points; Organization, 5 points;
Grammar, 4 points; Diction, 4 points;
Spelling, Capitalization and Punctuation, 2 points

It is worth noting that with this adjustment, Content and Organization now carried a greater weight, each given one more point while two points were deducted from the category of Mechanics (i.e., Spelling, Capitalization, and Punctuation). This change was significant in that it showed high school English teachers that more weight—and thus more emphasis—should be placed on the content and organization of students' compositions. This definitely was a change in the right direction, and undoubtedly had its positive backwash in teaching.

This scoring criterion has been adopted for marking examinees' compositions since 1993. To provide raters with more specific guidelines in marking examinees' compositions, the CEEC further developed the following rating scale (Chen et al. 1992, 1993), detailing the five components above:

Table 2: CEEC rating scale for marking examinee scripts

components	score range	descriptors
Content	5-4	**Excellent to very good:** well-stated thesis related to the assigned topic with relevant, substantive, and detailed supports
	3	**Good to average:** limitedly-developed or vague thesis with irrelevant statements
	2-1	**Fair to poor:** poorly-developed or obscured thesis; too much repetition of limited relevant sentences
	0	**Very poor:** not pertinent; or no written products (if this stands, all the other features are counted as "0")
Organization	5-4	**Excellent to very good:** well-organized structure with beginning, development, and ending; effective transition with logical sequencing and coherence
	3	**Good to average:** loosely-organized structure with imbalanced beginning, development, and ending; less effective transition that obvious affects logical sequencing and coherence
	2-1	**Fair to poor:** choppy ideas scattering without logical sequencing and coherence
	0	**Very poor:** no organization, no sequencing and coherence; or not pertinent
Grammar	4	**Excellent to very good:** well-structured sentences with variety; appropriate rhetoric; few grammatical errors
	3	**Good to average:** less well-structured sentence with some errors of tense, agreement, etc.; but meaning seldom obscured
	2-1	**Fair to poor:** major errors of conjunctions, fragments, or ill-structured sentences that make meaning confused or obscured
	0	**Very poor:** being dominated by errors that blocks communication
Diction	4	**Excellent to very good:** specific and effective wording; idiomatic and no spelling error
	3	**Good to average:** dull and repeated wording; occasional errors of word/idiom form, choice, usage but meaning not obscured
	2-1	**Fair to poor:** inappropriate wording; frequent spelling errors; meaning confused or obscured
	0	**Very poor:** some relevant words found, but meaning incomprehensible
Mechanics	2	**Excellent to very good:** no errors of format, punctuation, or capitalization
	1	**Fair to poor:** limited errors of format, punctuation, or capitalization, but meaning not obscured
	0	**Very poor:** too many errors of format, punctuation, or capitalization; violating basic conventions of writing

This analytic rating scale, adapted from Liu (2005, p. 98) with slight modification, has been adopted for marking examinee scripts in the GSAET, the DRET as well as the CEWPT over the years, and all the raters recruited for marking the scripts have to demonstrate their familiarity with the scale before the formal marking begins.

Here is a brief description of the standard CEEC marking procedures. Every

year, professors and lecturers with experience of teaching writing and marking written work are recruited by the CEEC. Depending on the number of examinees, the raters may range in number from 120 to 160 and are divided into groups, each with an experienced leader. Before the marking officially begins, the leaders of each group spend a whole day carefully examining a set of randomly selected examinee scripts (approximately1,500 in total), from which about 50 are to be chosen to serve as benchmark scripts. After careful discussion, the benchmark scripts are each assigned a proper score (out of a total of 20 points) based on the scale. These scripts are then divided into two sets, one for the training session and the other for trial marking by the raters. In each set, examinee scripts cover the full range of the scale. Then on the first day of the formal marking, each rater is presented with a booklet containing the task, the analytic scale, and the benchmark scripts. The leader first asks his/her group members to quickly go over the task and the scale and then proceed to examine closely the first set of benchmark scripts. The leader discusses the scripts with the members using the scale, explaining, in depth—if necessary, why a particular score is assigned to the script. After that, each group member begins to trial mark the second set of scripts on their own to demonstrate their familiarity with the use of the scale. The group leader then double-checks whether there are discrepancies in terms of the scores awarded to the scripts. If there is a mismatch, further discussion or explanation ensues. Only when there is complete agreement on the benchmark scripts can the official marking begin. This helps ensure inter-rater reliability as each examinee script is scored twice independently by different markers from different groups. To ensure fairness, scripts are submitted to a third, more senior leader if the scores awarded by the two markers show great discrepancy (i.e., 6 points or higher). It is worth mentioning here that due to the large number of examinee scripts to be marked and the limited time allotted for

marking, raters are allowed to use the holistic scale below (again, adapted from Liu 2005 with slight modification) after they have demonstrated reliability in marking an initial set of 200 examinee scripts.

Table 3: CEEC holistic scale for marking examinee scripts

score range	general comments on scripts
19-20	excellent
15-18	excellent-very good
10-14	good-average
5-9	fair-poor
0-4	very poor

The marking of examinee scripts was traditionally done in the paper method. However, to accommodate the growing test population and to facilitate the marking of examinee scripts, the CEEC began a series of research on the feasibility of marking examinee scripts on-screen (Chiu, et al, 2002; Lee, et al. 2004; Liu, et al. 2006). After two years of research, on-screen marking was first adopted in 2001 for marking examinee scripts in the CEEC English Writing Proficiency Test. In 2002 and 2003, this new method was employed for marking examinee scripts in the make-up exams of the GSAET. On-screen marking was again used in the 2008 CEEC English Writing Proficiency Test, and after these trials, the CEEC announced that on-screen marking would replace the conventional paper practice beginning with the 2011 GSAET.

On-screen marking has several advantages over the traditional paper-based mode. It significantly reduces the time and manpower needed for handling examinee scripts and for double-checking and recording examinees' scores. More importantly, on-screen marking makes it easier to observe raters' behaviors and monitor their consistency in marking, thus helping to ensure reliability in marking (Chiu, et al, 2002).

Take, for instance, the 2001 CEEC English Writing Proficiency Test. Examinees' hand-written scripts were first scanned, and digital images of the scripts were then displayed on screen for the markers to assign a proper score based on the rating scale. A total of 15,000 high school seniors took the test, which consisted of three forms. On-screen marking was adopted for rating examinees' responses in Form C, which comprised of three sentence-level tasks: (a) sentence-combining (10 questions), (b) sentence-making (6 questions), and (c) short question-and-answer (5 questions). 41 raters were recruited for the marking, and the same procedures for the traditional paper marking mode were followed, i.e., they were first presented with the rating scale for each task along with the benchmark scripts and were required to familiarize themselves with the scales. Before the formal marking started, the raters were further required to trial-mark another set to ensure marker reliability. The marking began on October 27 and ended on November 4, and generally speaking, the raters felt quite at ease with on-screen marking, though some senior raters specifically mentioned tiredness as a result of reading over an extended period of time on-screen. Some also suggested that improvement needed to be made in terms of screen resolution, script legibility, scrolling, system response time, seating arrangement, and lighting conditions (Chiu, et al, 2002). Improvements based on these suggestions have been made over the years to help ensure comfort as well as reliability in the marking of examinee scripts.

As mentioned above, on-screen marking was again adopted in marking examinee scripts in the 2008 CEEC English Writing Proficiency Test, which consisted of four tasks: (a) sentence-combining (5 questions; 4 points each), (b) translation in the cloze format (5 questions; 4 points each), (c) short letter-writing (20 points), and (d) essay writing (40 points), as shown below (Chang, et al. 2008):

Sentence-combining:

a. The flight was delayed.

b. There was a big storm. (with *because of*)

Translation in the cloze format:

1. 你喜歡吃什麼樣的食物？ Do you eat raw fish? Cheese? Many people prefer to eat food that they are familiar with. 2. 例如日本人喜歡吃生馬肉， but few Americans would want to taste it. 3. 有些人為了宗教的原因而不吃特定的食物。 For instance, Hindus do not eat beef because cows are considered sacred.

4. 然而，有時候我們必須改變飲食習慣。 If we travel to a new place with a different culture, our favorite food ingredients may not be available to us. 5. 我自己就有一個很慘痛的經驗 when I was traveling in Pakistan. For two days, I couldn't find anything I liked to eat.

Short letter-writing:

假設你是 Chris，因故將有一個星期無法到學校上課，請依下面的格式寫一封英文簡　函向陳老師請假，並簡要說明請假的理由。(Write a short letter to Ms. Chen on behalf of a student named Chris to request leave for a week and briefly explain why.)

03/21/2008

Dear Ms. Chen,

Essay writing:

台灣進行教育改革(educational reform)已有多年歷史,各界對其成效有褒有貶。請以一個高中學生的立場,點出現行教育制度最有成效之處或最嚴重的問題。英文作文須以"As a high school student, I think that the greatest success in Taiwan's educational reform is..."或以"As a high school student, I think that the most serious educational problem in Taiwan is..."開頭,並以自己的經驗舉例說明。

(From the standpoint of a high school student, write an essay of no less than 120 words on the topic of *Taiwan's Educational Reform*. You can either begin your essay with "As a high school student, I think that the greatest success in Taiwan's educational reform is..." or "As a high school student, I think that the most serious educational problem in Taiwan is..." Support your argument with concrete examples.)

For the marking of examinee scripts, 10 high school English teachers were recruited along with 36 college professors and lecturers. The purpose was to investigate whether or not high school English teachers would be as reliable as the professors and lecturers in marking the scripts. Each rater was assigned 192 to 262 scripts to mark, with high school teachers specifically assigned to mark scripts involving tasks (a) and (b). Ten professors and lecturers were also assigned to mark these two tasks for comparison in marker reliability. To further compare college-level and high-school level raters in terms of marking reliability, five high school teachers were later requested to mark 100 samples (3 marking task (c) (short letter-writing) and 2 marking task (d) (essay writing)).

As shown in Table 4, there was consistency in the marking of examinee scripts between the college professors and the high school English teachers, with the

Pearson product-moment correlation coefficients ranging between 0.856 and 0.949 ($p < 0.001$).

Table 4: Pearson *r* correlation between the two groups of raters in marking the four tasks

Task	*r*
Sentence-combing (N=483)	.949***
Cloze translation (N=356)	.934***
Short letter-writing (N=300)	.856***
Essay writing (N=200)	.916***

***p<0.001

Table 5 shows that except for task (a), the scores awarded by the high school English teachers to tasks (b), (c), and (d) were slightly higher than those awarded by the college professors, suggesting that the former tended to be more lenient than the latter in marking. In terms of effect size (Cohen's *d),* however, differences were found in the scores awarded by the two groups. According to Cohen (1988), when the effect size ranges between 0.2 and 0.5, small difference can be inferred. This means that slight difference existed between the marking of tasks (b) and (c) among the two groups, though no statistically significant difference could be inferred in the scores awarded to tasks (a) and (d) by the two groups (effect size <0.2).

Table 5: Results of t-test for the marking of the four tasks by the two groups of raters

Task	Rater	Mean	SD	*t*	p-value	Cohen's *d*
Sentence-combing (N=483)	College	14.547	3.865	1.346	0.179	0.004
	H.S.	14.470	3.964			
Cloze translation (N=356)	College	12.806	4.908	-8.308	0.000	0.403
	H.S.	13.584	4.633			
Short letter-writing (N=300)	College	8.652	3.881	-6.188	0.000	0.337
	H.S.	9.373	3.550			
Essay writing (N=200)	College	7.483	5.511	-0.616	0.539	0.004
	H.S.	7.535	5.654			

*College=college professors/lecturers; H.S.=high school English teachers

From the analysis above, it seems feasible for the CEEC to recruit high school English teachers for the marking of examinee scripts in the college entrance exams and the CEWPT. Rater training is definitely a must to better prepare the teachers to become qualified raters. Moreover, as slight difference was found between the two groups with regard to the marking of tasks involving cloze translation and short letter-writing, further research needs to be conducted to explore whether the tasks themselves are a major factor for the difference.

III. Conclusion

Ever since its founding in 1989, the College Entrance Examination Center has conducted extensive research in test development to better serve the stake-holders. In this paper, tasks for assessing Taiwanese high school graduates' English writing proficiency have been presented and their backwash on teaching discussed. Over the

years, the CEEC has also spared no efforts to ensure fair and reliable marking of examinee scripts. The rating scales, rater training procedures and the adoption of on-screen marking in recent years all point to this continuous effort. However, to relieve the tremendous pressure the CEEC faces every year having to complete the marking of examinee scripts within a limited period of time, the Center needs to further promote CEWPT—especially when high school English teachers have been shown to be able to help shoulder the responsibility of marking examinee scripts—so that discouraging and embarrassing scores (as shown in Table 1) will become a thing of the past. Moreover, as on-screen marking has now replaced the conventional paper mode for the rating of examinee papers, more research can be conducted to ensure comfort (e.g. how to improve the physical environments for marking the test papers), convenience (e.g. is it possible for the rater to do the marking at home) and, most important, reliability in marking (for example, reliability regarding the marking of the same script through the two modes by different raters or by the same raters but at different times). The addition of a fourth marker to adjudicate in the marking procedures in the 2011 GSAET is indeed a step in the right direction, but apparently there is much more to be done for the CEEC to fulfill its mission of serving the general public.

References

Bailey, K. M. (1998). *Learning about language assessment: Dilemmas, decisions, and directions.* Boston: Heinle & Heinle.

Chang, W. C. (1999a). Graded English writing proficiency test: A proposal. *In reference booklet for the seminar on graded tests of English language abilities*. Taipei: College Entrance Examination Center.

Chang, W. C. (1999b). *An analysis of the English test items of the 1998 scholastic achievement test.* Taipei: College Entrance Examination Center.

Chang, W. C. (2000). Test of written English in joint college entrance exam in Taiwan: Development and innovation. *Proceedings of the Third International Conference on English Language Testing in Asia*. Hong Kong: Hong Kong Examinations Authority.

Chang, W. C., et al. (2008). *CEEC English writing proficiency test: Project report*. Taipei: College Entrance Examination Center.

Chen, K., et al. (1992). *Item writing and scoring of JCEE English compositions*. Taipei: College Entrance Examination Center.

Chen, K., et al. (1993). *Scoring guide for marking JCEE English compositions.* Taipei: College Entrance Examination Center.

Chiu, M. C., Wen, J. S., & Liu, C. K. (2002). *A report on the 2002 cross-strait conference on College Entrance Examinations, appendix four: Computer-assisted scoring—A case study on English writing ability test.* Taipei: College Entrance Examination Center.

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (*2nd ed.*). Hillsdale, NJ: Erlbaum.

Cohen, J. (1992). A power primer. *Psychological Bulletin, 112* (1), 155–159.

Heaton, J. B. (1988). *Writing English language tests* (*New ed.*). New York: Prentice Hall.

Heaton, J. B. (1990). *Classroom testing.* New York: Longman.

Huang, T. S. (1994). *A qualitative analysis of the JCEE English test*. Taipei: The Crane Publishing Co., Limited.

Hughes, A. (1989). *Testing for language teachers*. Cambridge: Cambridge University Press.

Lee, M. Y., et al. (2004). *A report on CEEC's computer-assisted scoring project in 2003*. Taipei: College Entrance Examination Center.

Liu, C. K. (2005). Holistic scoring based on analytic features: Can holistic scoring be a constant winner? *Proceedings of the 8th Academic Forum on English Language Testing in Asia: Assessment For Learning.* Hong Kong: Hong Kong Examinations and Assessment Authority.

Liu, C. K., & Yao, H. L. (2006). The research and development of computer-assisted scoring system at CEEC. *Bulletin of testing and assessment. 1*, 49-72.

Weir, J. B. (1988). *Understanding and developing language tests*. New York: Prentice Hall.