科學教育學刊
2013, 第二十一卷第二期, 189-213
DOI: 10.6173/CJSE.2013.2102.03

Chinese Journal of Science Education
2013, 21(2), 189-213

高級中學自然科學術性向測驗編製

侯雅齡
國立屏東教育大學　特殊教育學系

摘要

科學資優是相當受到重視的學術性向資優領域，臺灣許多高中都設有科學資優班，以提供科學資優學生適性的教育服務。科學資優學生的鑑定方式之一，乃需要透過標準化的科學性向測驗以瞭解學生是否具有優異的科學潛能。本研究目的在編製適用於高中學生的自然科學術性向測驗，以作為鑑定科學資賦優異學生的工具。試題以科學思考與推理能力(scientific thinking and reasoning)的評估為核心，區分為物理、化學、生物、地科四範疇，所有題目皆經過嚴格的內容審查，並以全國高一學生為母群，抽取1,878名學生進行預試。正式測驗計有35題，其內部一致性Cronbach's α信度為 .813，間隔三週的重測信度為 .898，與「國中基本學力測驗的自然科學成就」的相關為 .782。以差異試題功能(differential item functioning)檢驗男女生在各試題的表現，並未有嚴重性別偏失的試題。最後以1,033位高一生為常模樣本，建立百分等級及常態化*T*分數常模。

關鍵詞：自然科學、測驗編製、資優學生、題目反應理論

壹、緒論

2009年11月修正公布之〈特殊教育法〉，在第十條關於特殊教育之實施，除了既有的學前教育與國民教育階段外，新增列了高級中等教育階段。此舉乃向上落實國家對不同智能、性向的學生，應給予適性、適切教育之理念。資賦優異學生相較於一般學生，除了較具學習潛能外，成就表現也相對的較為優異，若能提供他們充實學習之機會，或是提升他們面對挑戰之經驗，將有助於他們將潛在的能力充分展現出來，成為國家建設與經濟發展的人力基礎。

目前我國對於自然科學領域表現優異的高中學生，有「數理資優班」與「科學班」兩種安置班別。其中，國內為數較多的數理資優班乃援〈特殊教育法〉設立，在安置型態上，不同於國民中學以下階段，資優教育不能採取集中編班的方式，在考量高級中學階段學生，其興趣與能力已有所分化，故允許將有科學性向或能力的學生編入相同班級，給予整體性的課程規劃與活動安排，以滿足資賦優異學生學習上的特殊需求，期冀裨益於人才的拔尖。至於資優學生的篩選與鑑定，教育部另訂有〈身心障礙及資賦優異

(投稿日期：民國101年7月3日，修訂日期：民國101年10月8日，接受日期：民國101年10月25日)

06-侯雅齡.indd 189　2013/7/12 上午 11:46:31

學生鑑定辦法〉，其第十六條明訂對學術領域優異學生的鑑定標準如下：

> 一、前項任一領域學術性向或成就測驗得分在平均數正二個標準差或百分等級九十七以上，並經專家學者、指導教師或家長觀察推薦，及檢附專長學科學習特質與表現卓越或傑出等之具體資料。
>
> 二、參加政府機關或學術研究機構舉辦之國際性或全國性有關學科競賽或展覽活動表現特別優異，獲前三等獎項。
>
> 三、參加學術研究單位長期輔導之有關學科研習活動，成就特別優異，經主辦單位推薦。
>
> 四、獨立研究成果優異並刊載於學術性刊物，經專家學者或指導教師推薦，並檢附具體資料。

2007年教育部公告的資賦優異學生鑑定安置流程，更明確制訂出多元化、多階段的鑑定方式，並要求用來篩選與鑑定資賦優異學生的學術性向測驗及成就測驗，必需選用標準化評量工具。五年來，在成就測驗部分，全國皆參酌國中基本學力測驗的相關學科測驗成績與百分等級常模，較無工具不足的問題；但是在性向測驗的選用上，就顯得捉襟見肘，究其可能原因有二：首先，為符合鑑定標準，資優生的甄別工具須建有全國性的常模，以致測驗的編製耗時又所費不貲；再者，資優班榮耀標記產生的吸引力，使得國內用以作為資優鑑定的測驗，其內容的保密性普遍不佳(教育部，2008b)，為確保鑑定效度，測驗能重複使用的次數有限，堪用時間也短。Millman與Arter (1984)認為，若測驗經常要進行、需求許多複本測驗或客觀環境要求特定的測驗等等，就應考慮發展題庫(item bank)。資優學生鑑定是一個週而復始的工作，又有試題外洩的疑慮，因此，研究者以為，建立性向測驗題庫是值得嘗試的作法，而且只要題庫夠大，試題安全的顧慮就可以降低(Vale, 2006)。目前國內使用的「短效型」資優鑑定工具，也可能會因為安全考量，刻意避免雷同或重複的試題，而造成試題難度的提高，若有充足的題庫，也有助於確保測驗品質。本研究主要的目的在於編製高級中學自然科學術性向測驗，以評估學生在自然科學領域的學習潛能，但也考量國內資優鑑定工具的殷切需求，在編製過程中，採用題庫建置的概念進行，希望透過本研究先建立起已校準試題(calibrated items)，讓未來得以此為基礎，逐步擴充試題數量，建立更完備高中科學性向測驗題庫。

一、學術性向測驗的內涵

學術領域資賦優異學生的鑑定，主要依據學術性向測驗或成就測驗得分，一般認為，成就測驗在測量學生於學校習得的知識與技能，而性向測驗則較不受學校學習經驗限制，較著重預測個人未來可能發展的潛在能力(potentiality)，也就是說，性向測驗有助於瞭解個人若有機會學習或接受訓練時，可能達到的程度，而非個人已學得的知識經驗或對某些作業熟稔的程度。

目前常用的性向測驗可以粗分為普通性向測驗與特殊性向測驗，兩者的差別在於測驗內容：普通性向測驗，多用來測量語文、數量、空間知覺、抽象思考、邏輯推理等方面之潛能，類似一般智力測驗；而特殊性向測驗多為單項測驗，用以測量受試者某方面的特殊潛能，如音樂、美術、數學、科學、機械、文書等。

學術性向測驗屬於特殊性向測驗，Cronbach提出能力測驗光譜(spectrum)說明成就測驗與性向測驗間的關係，以及不同類型性向測驗之間的關係(引自Linn & Miller, 2005／王振世、何秀珠、曾文松、彭文松譯，2009，頁459)，他認為成就測驗和性向測驗均屬於能力測驗，只是在測驗內容的性質上有連續性的差異，根據表1的能力測驗光譜，水準A是內容導向的成就測驗，測驗內容針對特定課程內容而設計；而水準E則是文化導向的非語言性向測驗，其測驗內容為較不受學校經驗影響的學習類型。由水準A到水準E所測量的學習內容，越來越不依賴特定的學習經驗，成就測驗的等級落在水準A和水準B，而性向測驗則是屬於水準C至水準E，其中學術性向測驗所測量的內容與學校經驗較接近，故在等級上比較接近成就測驗。本研究所欲發展的自然科性向測驗，目的在預測學生未來自然科學的學習之成效，因此預測基準乃建構於學生在自然科學所具備的學識背景。

表2是國內常用於高中學生的性向測驗，從測驗內容可知，多數是屬於普通性向測驗，僅有「高級中學數學及自然性向測驗」能針對學生在自然學科領域性向進行瞭解，李秉書(2003)以臺灣高中學生為對象，探討物理的學業成就與多因素性向測驗的關連性，發現兩者之間並無必然關係，也就是說，若以普通性向測驗作為自然科學資優學生性向評估，並不十分適切。Zaaiman, Flier與Thijs (2000)也認為，若要預測學生未來學習之成效，預測基準必須建構於學生在學科已具備的學識背景之上。本研究聚焦於自然科學領域，編製的自然科性向測驗乃為協助資優學生的鑑定，希望瞭解受試者在自然科方面的能力，以提供其適性的教育環境，助其潛能之開展。所以，本測驗的內容向度乃依據教育部2008年公告之《普通高級中學課程綱要》內涵，作為構思命題架構的考量，初分為物理、化學、生物、地球科學及生活科技(教育部，2008a)。

至於能力向度規劃，則廣泛參考國外常見的學術性向測驗，包含：美國大學測

表1：Cronbach根據測量的學習類型所排之能力測驗光譜

水準	一般測驗類型	測量的學習類型
A	內容導向的成就測驗 如：成就與能力測驗	課程中的學科知識，如社會、英文、數學與自然。
B	一般教育發展測驗 如：愛荷華教育發展測驗 (Iowa Test of Educational Development, ITED)	許多課程共有的基本技能與複雜的學習結果，如原則的應用與資料的解釋。
C	學校導向性向測驗 如：預備學術評量測驗 (preliminary scholastic assessment tests)	類似學校所學的語文及數量的問題解決能力與寫作技巧，如字彙、閱讀與數學推理。
D	文化導向的語文性向測驗 如：認知能力測驗、語文綜合測驗與數量綜合測驗等	來自一般文化的語文、數字及一般問題解決能力，而非共同的學校經驗。
E	文化導向的非語文性向測驗 如：認知能力測驗、非語文綜合測驗	以圖像類比、圖像分類及其他與學校經驗無關的非語文作業為基礎之抽象推理能力。

資料來源：引自Linn, R. L., & Miller, M. D. (2009)。教育測驗與評量(*Measurement and assessment in teaching*; 9th ed.；王振世、何秀珠、曾文松與彭文松譯，頁459)。臺北市：雙葉書廊。(原作出版於2005年)

表2：國內目前適用於高中學生的性向測驗

測驗名稱	作者	出版年	內容		測驗對象
多因素性向測驗	路君約、盧欽銘與歐滄和	1994	1.語文推理 2.數學推理 3.機械推理 4.空間關係	5.抽象推理 6.錯別字 7.文法與措辭 8.知覺速度與確度	國二到高三學生
中學多元性向測驗	毛國楠與盧雪梅	2003	1.語文推理 2.數學能力 3.科學推理 4.知覺速度與確度	5.空間關係 6.抽象推理 7.字詞釋意 8.文法修辭	八年級至高二
多向度性向測驗	歐滄和與路君約	2003	1.語文推理 2.機械推理 3.電腦能力 4.數學推理	5.抽象推理 6.電學知識 7.文句重組 8.資料核對	八年級至高二
高級中學數學及自然性向測驗	郭靜姿、吳淑敏與黃玉蓮	2006	1.數學科性向測驗 (1) 數量關係 (2) 圖形空間 (3) 推理能力	2.自然科學性向測驗 (1) 物理 (2) 化學 (3) 生物 (4) 地球科學	高一學生
高一性向測驗	簡茂發等	1994	1.語文詞義測驗 2.數學能力測驗 3.語文推理測驗	4.機械推理測驗 5.空間關係測驗	高一學生
高中系列學業性向測驗	黃堅厚與盧欽銘	1986	1.語文推理	2.數量比較	高一至高三學生

驗(American College Test, ACT)、全國性教育進展評量(National Assessment Educational Progress, NAEP)、瑞典學術性向測驗(Swedish Scholastic Aptitude Test, SweSAT; Aberg-Bengtsson, 2005; Stage & Ögren, 2004)、加拿大學校成就指標計畫(School Achievement Indicators Program, SAIP; Council of Ministers of Education, 2005)與泛加拿大評量計畫(Pan-Canadian Assessment Program, PCAP; Education Quality and Accountability Office, 2008)等，在自然科學部分所評量的能力。

由表3所彙整之測驗的能力向度評量焦點來看，所有測驗都包含了對高層次思考與推理能力的測量，本測驗主要用途在於篩選有自然科學性向的國中畢業生就讀高中數理資優班，所以測驗的重點在於學生能否利用已有的科學知識和概念進行更高層次的思考與推理。對於科學的思考與推理，有偏重於詮釋觀察現象之間關係，強調經驗—歸納(empirical-inductive)者，如Zimmerman (2005)重視根據以往的學習，對觀察現象提出假設、評估與驗證；亦有偏重於學習理論性概念與概念改變，強調假設—演繹(hypothetical-deductive)者，如Hogan, Nastasi與Pressley (1999)認為科學推理是抽象邏輯的假設演繹與背後因果的推測。然而，一般認為至少具有以下內涵(Dunbar & Fugelsang, 2005; Thagard, 2005)：(一)假說檢驗：由蒐集到的事實證據去評價命題；(二)問題解決：有計畫與步驟的尋找關係，進而建構因果模式；(三)類比推理：由已知的發現或理解，連結到欲理解的

表3：國外常見的學術性向測驗

測驗名稱	能力向度規劃
美國大學測驗	1.資料的陳述 2.研究摘要 3.爭論的辨別
全國性教育進展評量	1.概念理解 2.科學探究 3.實際推理
瑞典學術性向測驗	1.訊息理解 2.圖表分析
加拿大學校成就指標計畫與泛加拿大評量計畫	1.自然科學在科技與社會議題的應用 2.概念性知識 3.科學知識應用

新事件中；(四)歸納與演繹推理：在仔細觀察後整理出影響的原理，進而形成理論，而演繹則是由一個前提推演出後續的結論。

Lawson (1994)對科學推理的定義，周延地含括「經驗—歸納」與「假設—演繹」，也接近一般性對科學推理的看法，因此，本測驗以「經驗—歸納推理」與「假設—演繹推理」能力，作為核心評估的能力，其中，「經驗—歸納推理」意指蒐集事實，依據事實去找共通可信的事項或概化原則。也就是說，根據以往的學習，對觀察現象進行詮釋。其中包含兩大技巧為：事證(evidence)蒐集，以及由現有資料去進行假設與驗證。而「假設—演繹推理」則是將概化的原則，或真實解析應用到具體事項上，也就是說，面對眼前事物或訊息，輔以過去經驗及知識，綜合出新的事物或資訊。

另外，也考慮資優生善於面對複雜性高的問題之特質，因此所有命題原則為：由已有資料(文字、數據或圖表)找出其特性、規則或關係，以尋求解決問題的方法，是間接性知識的獲得。因此在所有試題中，學生的解題步驟設定為兩步驟或三步驟。

二、題庫擴充時的等化應用

題庫，顧名思義是一堆試題的集合，但是一個素質好的題庫，必需在心理計量的基礎上，經過嚴謹的編製歷程、試題校準(item calibration)程序、分析、歸類與評鑑後，貯存起來供特殊使用的試題集合(余民寧，2009)。一般測驗的發展多依據古典測驗理論(Classical Test Theory, CTT)，但是以CTT所發展的測驗，其試題難度與鑑別度參數，皆受到受試群體能力差異的影響，雖然試題相同，但使用於不同能力的群體時，試題的難度與鑑別度就會改變，使參數的參考值受限。而試題反應理論(Item Response Theory, IRT)所獲得的試題參數與所測量之群體相互獨立，不會因受試群體之不同，而改變其難度、鑑別度等參數，此一特性為嚴謹題庫建置的良好基礎。再者，由於每一個試題對測驗訊息函數的貢獻可以獨立決定，將試題訊息作累加就構成測驗的訊息，試題組卷時可以依據目的，在理想的測驗訊息曲線下，由題庫中選擇適切的試題來建構貼近理想的測驗訊息(余民寧，2009；吳裕益，出版中；DeMars, 2010; Embretson & Reise, 2000; Miller, Vandome, & McBrewster, 2009)。

題庫中的試題必須建立在一個共同的量尺(common scale)之上，如此試題方能相互比較與使用(Kolen & Brennan, 2004)，IRT提供測驗等化(equating)方式，讓接受不同測驗的不同考生，其能力可以相互比較，也讓多份測驗的試題參數估計值，放置在一個共同單位的量尺，讓不同測驗的難度水準可以相互比較，也讓題庫規模可以擴充(Dorans, Pommerich, & Holland, 2007; Vale, 1986)。

不過，要進行測驗等化，必須假定所要等化的測驗是單一向度，並蒐集適合等化研究的資料。例如，要有同時參加幾式測驗的共同受試組(repeated subjects)，或是兩式測驗必須有共同試題(common item)。綜合來說，可用來進行等化研究的資料蒐集設計可以歸納出以下七種(吳裕益，出版中；Angoff, 1984; Dorans et al., 2007; Kolen & Brennan, 2004; Lord, 1980)。

(一)單組設計

相同受試皆接受兩式需要銜接(linking)的測驗，且兩式測驗需在同一天進行，以可避免兩式測驗難度水準差異，與參加每一式測驗的受試組，能力水準差異相混淆的問題，但是受試的疲勞因素與重複練習因素，不免會影響第二式測驗的參數估計結果。

(二)對抗平衡隨機組設計

考慮到單組設計，受試接受測驗之先後順序會影響得分。對抗平衡隨機組設計，將所有受試分為兩個隨機的半組，而且兩半組依對抗平衡順序同時參加兩式測驗，也就是一組先參加其中一式測驗，而另一組則先參加另一式測驗。

(三)等組設計

等組設計是單組設計及對抗平衡隨機組設計的簡單替代設計，此種設計不需要讓每位受試參加每一式測驗。要等化的兩式測驗分別用來測驗兩個隨機組中的一組，為減少兩組受試在所要測量能力的差異，需使用大樣本來降低等化過程的偏差。

(四)定錨測驗(或共同測驗)隨機組設計

定錨測驗意指共同測驗，不同測驗除了各自施測不同群體外，還有一個共同測驗用以同時測量兩組受試，定錨測驗的分數可提供不同測驗共同的參照點，以減少來自於兩組受試能力差異所造成的等化偏失(equating bias)。

(五)定錨測驗不等組設計

相較與隨機組設計，此種設計蒐集資料的受試群體為不等組或自然產生的組別，比較可能會有系統的差異，尚無統計方法可以提供完全適切的調整。

(六)分測驗預先銜接等化設計

上述五種資料蒐集的設計需要以未使用過的完整測驗(intact tests)來測量取自某些母群的受試樣本。相反的，分段預先等化的資料蒐集設計，是在Y測驗成為完整的測驗之前，就將Y式等化到X式測驗。有些測驗的內容是由數個時段所組成。某些段稱為「現用的」(operational)，其他稱之為「可變的」(variable)。現用段在所有試卷均相同，各現用段的得分加起來即為所報告的全測驗分數，各題本可變段的內容互不相同，其目的可以預試新試題或實驗新題型，可變段的得分不計算到全測驗分數內。

(七)試題預先等化設計

此一設計最適合用來產生一個已校準試題(calibrated items)的題庫。當有試題已得到共同量尺的IRT參數估計值，再根據已校準的試題建立未來使用的新測驗並且加以等化。

此種設計的必要條件是，在每一個校準分析(同時估計一組或多組試題的IRT參數之程序)時，每一受試組至少需要與其他一個受試組有某些共同試題，而且至少有一組試題必須是預先校準的(precalibrated)。在試題預先等化設計，受試對試題的反應可能會受試題排列序位不同，和兩次施測周遭試題間內容差異的影響。這些因素可能使預試時的試題變得比正式實施測驗時容易(或較難)以及使鑑別力更大(或較小)，造成等化的偏失。

本研究在綜合考量各種現實因素，以及冀求產生一組校準試題，以利未來試題規模擴充，故採用定錨測驗方式將通過內容與文化審查的所有試題，分成多份測驗，在測驗間納入共同試題進行水平等化(horizontal equating)，使能力估計值等值，且試題參數的估計值也在同一量尺。俟取得所有試題的鑑別度、難度和猜測度三參數後，考量施測目的在鑑別資優生，故以介於+1至+2的測驗訊息函數，選取適合的試題來組成所要的測驗。由於資優鑑定測驗需建置全國性常模，完成常模建置的正式測驗，很適合作為已校準試題，未來研究只要採用「試題預先等化設計」方式，即可利用本研究之已校準的試題，逐步透過共同試題使二卷的試題參數化為相同量尺單位，取得不受樣本影響的獨立試題參數，藉此不斷擴充題庫規模。

三、IRT取向之差異試題功能

一份可靠與有效的測驗，影響測驗結果的主要變異應聚焦於所欲測量的構念，而不受其他無關的變異的影響。如果受試者在試題的表現受到無關變異的影響，或者這些無關變異對於某些群體有系統性的影響，都會影響測驗的品質。當所屬次級群體有別，以致在回應試題時有不同的表現，那便顯示同一試題對兩個群體有不同的功能，而造成差異試題功能(Differential Item Functioning, DIF)現象(Holland & Wainer, 1993)。差異試題功能現象，除了由試題內容判斷外，最好的方式還是透過實徵研究(empirical study)實地探討各次級群體在每一試題的實際表現偏差(吳裕益，出版中)，從心理計量取向的觀點來看，DIF的定義為相同能力的人，若因為來自不同群體，以致在某一試題的答對機率有所不同，該試題便顯現出DIF (Angoff, 1993)。

目前國內、外發展較為嚴謹的測驗時，為求測驗內容對各個群體公平，多會以性別、城鄉、教育資源或生活經驗進行不同群體的比較，例如，大學入學測驗曾就教育資源、就讀學校、性別、類組進行分析(大學入學考試中心，2001，2005)，盧雪梅與毛國楠(2008)及余民寧與謝進昌(2006)對國中基本學力測驗試題進行不同性別、區域、城鄉或生活經驗的分析。

在「男理工，女人文」的印象下，一般認為自然科似乎是對男生有利的學科，且Han與Hoover (1994)及Zhang與Manon (2000)的研究，都發現性別差異與能力高低有交互作用的現象，能力低的群體中，女生的表現優於男生，能力高的群體，男生的表現優於女生，盧雪梅與毛國楠(2008)分析歷時5年的國中基測自然科測驗之性別DIF，結果發現DIF出現率約為9.0%，其中健康教育、化學和生物三科的DIF題有利女生居多，地球科學和物理的DIF題有利男生居多。本研究所發展的測驗，目的在瞭解學生在自然科的學習潛能，作為資優學生鑑定安置之用，因此，測驗對男女生群體的公平性如何，應該也要被仔細考量。

CTT理論以變異數分析、轉換試題難度法、相關法、因素分析法等進行的DIF檢驗，

都受到樣本依賴的限制；若取樣規模夠大時，以IRT為基礎衍生的方法相對受到推崇，IRT提供了「試題參數比較法」、「試題特徵曲線比較法」與「適合度比較法」三種方式來比較DIF，Hambleton與Swaminathan (1989)推薦使用試題參數比較法和試題特徵曲線比較法。這兩種方法本質上相同，都是在比較特徵曲線，只是兩種方法之操作性定義不同：

(一)試題參數比較法

此法類似多變項變異數分析法，考驗兩個群體在每一個試題的參數(a, b, c)是否達顯著差異，如果兩個試題特徵函數的參數值相同，答對該試題的正確率應該會相同，所以此項考驗的統計假設如下，當某一試題的虛無假設被推翻，即表示該試題具有DIF現象。但此方式必須要在大樣本之下才適用，且兩試題特徵曲線縱使無實質差異存在，也會有很顯著的試題參數值，以致卡方分配統計值有較高的偽陽性反應率(false-positive rate)容易產生誤判。

(二)試題特徵曲線比較法

此種方法在於比較不同群體，在相同試題所夾面積之差異情形，當面積差異越大，代表試題的DIF程度越大。理論上，試題參數的估計不受考生能力分佈的影響，不同群體考生在同一道試題所估計的參數經等化建立共同量尺後，其試題特徵曲線也會相同，因此，面積之差異也應該等於0。不過，當採用三參數的對數型模式時，必須假設c參數在兩群體皆相等，否則面積將變得無限大，以致顯著性考驗無法進行，另外，若樣本不夠大，能力值的範圍可能不夠寬廣，也容易導致錯誤的DIF結論。

在本研究中則採用試題參數比較法來進行性別的DIF研究。

貳、研究方法

一、研究參與者

本研究之樣本包含預試樣本、常模樣本與重測樣本，因為目前高中數理資優班的甄試，全國統一在八月中下旬進行，因此預試與常模樣本的施測時間，乃安排在各校甫開學即進行施測，以避免教學後所造成的干擾，影響測驗的效度。

(一)預試樣本

我國高中入學方式尚有明顯的能力取才，故預試樣本選取前，乃將各校依學生能力分層後再進行叢集取樣，不過學生能力分層方式並無具體可參酌之文獻，研究者參考九十八學年度各校錄取分數，將其依序排列後，再考量學校聲望，做一般化的分級，部分私立高中職採獨立招生方式，其錄取分數的參考價值不高，並未納入分級，也未在選取樣本之列，另外，也將全國高中區分成北、中、南區後隨機抽取學校之部分新生作為受試。

各層級學生人數的分佈，因考量到本研究之自然科性向測驗乃用於甄別資優生，試題偏難，為了在預試階段能取得學生合作，有最大動力作答，以利蒐集到有效的試題訊息作為未來選題組卷的依據，所以選取了較多的一、二級學校。由表4可知，預試樣本包含十二所學校的高一新生，十二所高中為：北一女中、臺中一中、臺南一中、成淵高中、惠文高中、家齊女中、南港高中、嘉義女中、屏東女中、中山高中、屏北高中、岡山高中。

所有試題分組成甲、乙、丙、丁四卷進行預試，由表5中可見，甲卷總樣本數為481人，有效樣本為478人，有效樣本比例為99.3%；乙卷總樣本數為456人，有效樣本為

表4：預試樣本分層

區域＼級別	第一級	第二級	第三級	第四級	第五級
北區	北一女中	成淵高中	南港高中	中山高中	
中區	臺中一中	惠文高中	嘉義女中		
南區	臺南一中	家齊女中	屏東女中	屏北高中	岡山高中

455人，有效樣本比例為99.7%；丙卷總樣本數為488人，有效樣本為485，有效樣本比例為99.3%；丁卷總樣本數為463，有效樣本為460，有效樣本比例為99.3%。整體看來，由於事前試題組卷時已考慮學生作答的負擔，再加上所有施測工作皆由兩位經過訓練的主試人員施測，嚴格執行標準化流程，故各試卷有效率高達99.3%以上。

(二)常模樣本

常模樣本的選擇乃參考預試樣本取樣模式，依據九十九學年度各校學生錄取分數，將各高中依次排序後，根據常態分配比例，與臺灣北、中、南、東各區學生數比例，共選取十四所高中(詳如表6)的高一新生參與常模建置。並於各高中開學後兩週內完成施測。

由表7可得知，總樣本數為1,082人，有效樣本為1,033人，有效樣本比例為95.4%。

(三)重測樣本

研究者分別在預試階段與正式階段進行重測，以利信度研究成果作交互驗證，預試階段在預試完成後二週，選取臺南一中與屏

表5：預試樣本

校名＼卷別狀況	甲卷		乙卷		丙卷		丁卷	
	施測人數	無效問卷	施測人數	無效問卷	施測人數	無效問卷	施測人數	無效問卷
1. 北一女中	41	0	41	0	41	0	41	0
2. 中山高中	39	0	40	0	37	0	22	3
3. 南港高中	37	1	37	0	36	1	35	0
4. 成淵高中	38	0	35	0	38	1	36	0
5. 臺中一中	44	2	44	1	44	0	42	0
6. 惠文高中	42	0	42	0	43	1	43	0
7. 家齊女中	39	0	42	0	42	0	42	0
8. 臺南一中	43	0	43	0	43	0	42	0
9. 嘉義女中	42	0	28	0	42	0	41	0
10.屏東女中	42	0	30	0	42	0	42	0
11.屏北高中	37	0	37	0	40	0	40	0
12.岡山高中	37	0	37	0	40	0	37	0
小計	481	3	456	1	488	3	463	3
總計	478		455		485		460	
有效卷比例	99.3%		99.7%		99.3%		99.3%	

表6：常模樣本分層

	第一級	第二級	第三級	第四級	第五級
北區	建國中學	景美女中	新店高中	安樂高中	
中區		臺中二中 忠明高中	員林高中 斗六高中		
南區		鳳山高中 臺南二中	左營高中	永仁高中	潮州高中
東區			羅東高中		

表7：常模樣本

狀況 / 校名	總樣本	
	施測人數	無效問卷
1. 安樂高中	53	4
2. 建國中學	90	1
3. 景美女中	77	0
4. 新店高中	84	1
5. 羅東高中	75	0
6. 臺中二中	72	1
7. 忠明高中	92	6
8. 員林高中	79	4
9. 斗六高中	80	2
10.永仁高中	73	7
11.臺南二中	81	4
12.鳳山高中	79	2
13.左營高中	74	12
14.潮州高中	73	5
小計	1,082	49
總計	1,033	
有效卷比例	95.4%	

東女中學生作為重測樣本；在正式階段，則於施測完成後三週，選取臺中二中與潮州高中學生進行重測，兩階段重測樣本數統計詳如表8。

二、測驗編製程序

本研究以編製自然科性向測驗為目的，考量到我國資優生鑑定對測驗工具的需求量大，故以題庫建置的概念來進行，資優生鑑定的試題，往往有試題外洩與保密性不佳的問題，必須有夠大的題庫才能因應，研究者採用試題反應理論進行測驗編製，期能取得所有試題的試題參數，讓未來只需要採用試題參數連結方式，即可擴充題庫。再者，本研究編製的測驗建有全國性常模，所以可獲得代表高一學生樣本通過每一試題比率的估計值，未來也只要以適當的樣本來施測舊測

表8：預試階段與正式階段重測樣本

	預試階段				正式階段
	甲卷	乙卷	丙卷	丁卷	正式卷
施測人數	85	73	85	84	139
有效問卷	84	72	82	82	126
有效卷比例	98.8%	98.6%	96.4%	97.6%	90.6%

驗與新試題，即可獲得所有試題的參數，並據以估計全國學生潛在特質水準的分配，有益於未來試題擴充的方便性。圖1為測驗編製過程，以下乃一一詳細敘述之：

(一)決定命題原則與架構

由文獻探討中已確定本自然科性向測驗，在Cronbach的能力測驗光譜中傾向於C級水準，其測驗內容應與未來學習內容有關聯，因此測驗命題架構乃含括學生科學學習領域，包含物理、化學、生物、地球科學與生活科技。而能力內涵以科學思考與推理能力(scientific thinking and reasoning)的評估為核心，以測驗學生「歸納推理」與「演繹推理」的高層次思考能力。為使題庫內的試題可以提供更豐富的訊息，提供給命題委員的命題紙，乃根據各能力向度的內涵，進一步細分為：事實蒐集、假設驗證、因果關係、推論預測、問題解決等細項。

為了使命題委員們瞭解本測驗的用途，並能順利進行測驗架構的討論及命題，研究者另編寫命題手冊，手冊內容包含有：

1.測驗目標、命題方向。
2.命題紙格式及填寫示例。
3.高中階段資優學生鑑定安置流程。
4.「自然與生活科技」學習領域之教材內容要領。
5.一般命題原則與選擇題命題原則。

(二)分領域命題

邀請具物理、化學、生物、地球科學及生活科技專長，且有多年指導學生參與科展經驗的五位國、高中教師參與第一階段命題工作，在審題階段，則再敦聘二位物理、化學專長之大學教授於概覽全部試題後，進行第二階段命題工作，使試題內容更加多元。五位中學教師包含二位高中教師，在考量到科學為男生的優勢學科，命題老師應有好的性別警覺意識，因此兩位高中老師乃分別邀請任教於二所女子高中的教師。所有教師皆有豐富的教學經驗，及優異的教學表現，並取得該領域碩士以上學歷。

研究者與命題委員之間透過多次命題會議，充分討論命題原則，並根據測驗向度規劃表，及各教師多年教學經驗，提出在自然科學領域中，學生所應具備的各種能力，包含歸納、演繹、觀察、推理、預測、驗證、問題解決等，並定義這些用語，以及其在「歸納推理」及「演繹推理」命題架構的理想歸屬。所有試題在於瞭解學生能否對科學方面知識融會貫通，而進一步進行因果推理或根據既有的科學定律去解釋日常現象，所以試題須以學生熟知的素材或背景知識為前提進行命題。因為研究者對於測驗編製專業有較多瞭解，命題初期也一一與命題教師的就試題設計方式做個別討論，協助命題教師充分掌握命題原則。

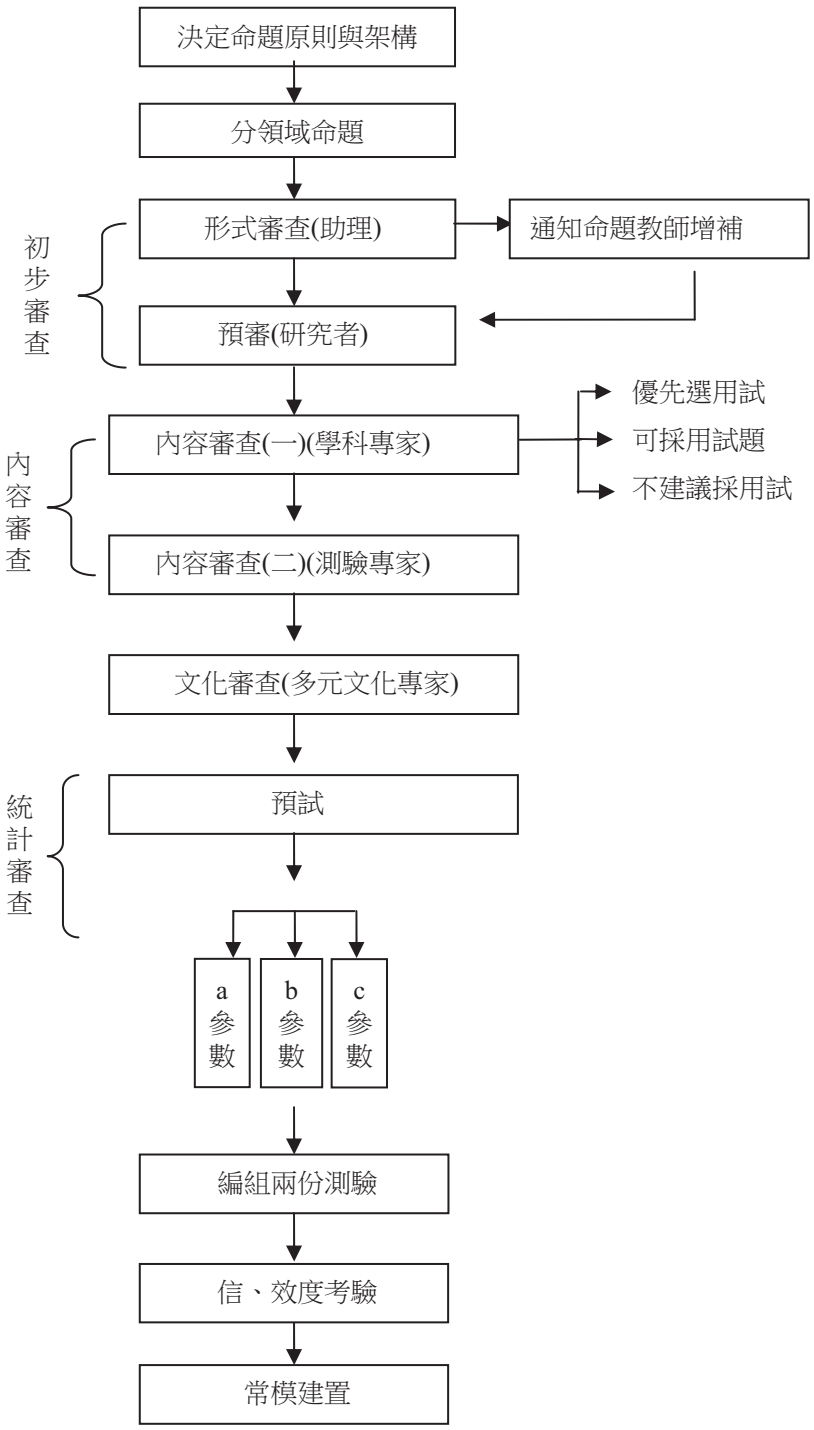

圖1：高級中學自然科學術性向測驗編製流程

(三)試題審查

在收題後，即進行試題審查，本測驗建構歷程的試題審查方式包含：初步審查、內容審查、文化審查及統計審查，以下則分別加以說明：

1.初步審查

此一階段審查目的在確認命題委員之命題是否符合形式要件及命題原則。總計回收試題包含：物理52題、化學36題、生物37題、地球科學23題、生活科技25題，合計共173題。全部的試題先由助理進行形式審查，確認試題符合形式要件，包含：領域、次領域是否明確、是否勾選能力要項(歸納推理、演繹推理)、解題步驟(兩步驟、三步驟)、預估難度值(易、中或難)、題幹主要出處與解題所需能力概述是否清楚。

完成形式審查後，則由研究者進行複審，檢視是否符合命題原則，再交由內容審查委員審查。其中在生活科技領域之試題，研究者依試題內容判斷各題偏向領域後，將其分送相近領域專家進行審題，在剔除明顯違反命題原則的試題後，交付內容審查的試題數為148題：物理66題、化學31題、生物31題、地球科學20題。

2.內容審查

內容審查專家包含生物、物理、化學、地球科學四個專業領域之大學教授與測驗領域的教授，審查結果說明如下：

(1)試題內容審查及修改

學科專家需對每一試題填寫「試題品質審查意見表」，分別對於：能力要項與題幹是否符合？評量目標與題幹是否符合？題意是否清楚明確？答案是否正確且為唯一解？一一作檢核，並對各試題的選項品質與誘答力進行評估；最後，在總評中勾選各試題的推薦程度，推薦程度區分為：優先選用試題、可採用試題、不建議採用試題。對於小幅修改即可採用的試題，則由審題委員對題幹或選項逕行修正。

表9是分別對各科試題推薦程度之統計。由表中可知物理有51題被推薦，化學則有22題被推薦，生物與地球科學被推薦試題分別為25題與17題。未受推薦的試題，以題意不夠清楚明確居多數，選項品質與誘答力不足居次。

(2)試題編製原則審查

剔除學科專家不建議採用試題後，再委由測驗專家對試題內容是否符合預擬的內容界定、試題的題幹是否簡潔明確，以及試題誘答是否有效等，進行審查與文句修正。此階段的審查保留了所有經內容審查後所建議採用的試題，但審查者對部分試題的題幹敘述與選項敘述做了文句修正，使試題的陳述更趨簡潔。

3.文化審查

為避免試題存在對次級文化的偏見，亦委請一位對次級群體文化差異敏感的學者

表9：學科專家對各科試題推薦程度之統計

類科 / 推薦程度	物理	化學	生物	地球科學	合計
優先選用題	39	10	15	9	73
可採用試題	12	12	10	8	42
不建議採用試題	15	9	6	3	33
總題數	66	31	31	20	148

進行試題的文化審查，審查重點在於：試題敘述是否有不當語文或情境，以及是否有文化、性別與人權等意識上之偏見，希望能避免不必要的文化等意識型態，以確定測驗的公平性。此階段的審查結果，115題試題皆未有任何對次級文化的偏見。

4.統計審查

由於試題以高層次思考、推理能力為核心，受試者作答時需要充分的時間與專注力，若一次施測過多試題易造成受試者疲累而影響效度，故乃將通過逐層審查後的試題分為甲、乙、丙、丁四卷(除丁卷為34題外，各卷皆為35題)進行預試，每卷施測時間為50分鐘，希望藉由學生實際反應資料來進行量化分析(quantitative analysis)，瞭解試題品質。

研究者在預試測驗組題時，也同步考量學科領域與試題難易度的分配，務使四份問卷在形式與作答時間接近。此外，因未來正式測驗選題組卷方式，將應用IRT之測驗發展的歷程，故須以logistic三參數模式取得所有試題的鑑別度參數(discrimination parameter, a參數)、難度參數(difficulty parameter, b參數)及猜測度參數(guessing parameter, c參數)。為了讓來自不同樣本、不同測驗卷所獲得之參數可以做獨立的比較，研究者在甲、乙、丙、丁卷中皆納入各學科領域不同難度的試題做為共同試題，期冀藉由等化方式取得不受樣本影響的獨立參數，由圖2可知甲、乙卷之間，乙、丙卷之間，以及丙、丁卷之間，都有共同試題，表10可進一步檢視物理、生物、化學與地科共同試題分佈，以及在各卷完整試卷中，共同試題如何安排在相近的題序。

三、資料處理方式

本研究使用SPSS 17.0和BILOG-MG 3.0統計軟體進行資料處理：

(一)以次數分配、描述統計方式先進行資料檢核，以免錯誤的資料輸入導致偏差的分析結果。

(二)進行試題等化及3PL分析，並在設定的目標訊息曲線下，選取適當的試題組卷。

(三)以Cronbach's α係數考驗測驗之內部一致性信度。

(四)以積差相關瞭解測驗與基本學力測驗之

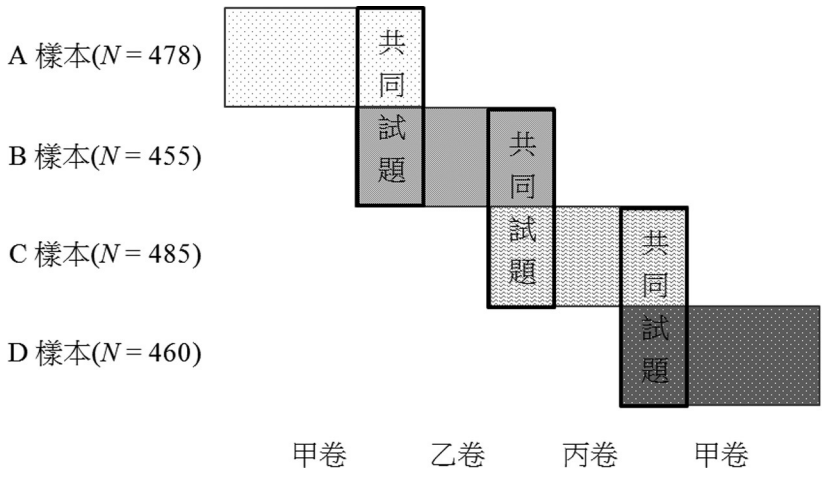

圖2：四份試卷共同試題配置

表10：四份試卷試題題號對照表

題庫	甲	乙	丙	丁	題庫	甲	乙	丙	丁	題庫	甲	乙	丙	丁	題庫	甲	乙	丙	丁
P_1	2				P_{30}			7		B_8		13			C_{12}		9	9	
P_2	6				P_{31}			11		B_9		19			C_{13}			14	
P_3	7				P_{32}			12		B_{10}		20			C_{14}			19	
P_4	11				P_{33}			16		B_{11}		25			C_{15}			26	
P_5	12				P_{34}			17		B_{12}		30	34		C_{16}			31	34
P_6	16				P_{35}			22		B_{13}		8	8		C_{17}			4	4
P_7	17				P_{36}			23		B_{14}			13		C_{18}				9
P_8	23				P_{37}			28		B_{15}			18		C_{19}				15
P_9	24				P_{38}			29		B_{16}			24		C_{20}				20
P_{10}	26				P_{39}			32	32	B_{17}			25		C_{21}				25
P_{11}	27				P_{40}			1	1	B_{18}			30	33	C_{22}				30
P_{12}	30				P_{41}				2	B_{19}			3	3	E_1	10			
P_{13}	31				P_{42}				6	B_{20}				8	E_2	15			
P_{14}	34				P_{43}				7	B_{21}				13	E_3	21			
P_{15}	35	33			P_{44}				11	B_{22}				14	E_4	22	22		
P_{16}	1	2			P_{45}				12	B_{23}				19	E_5	5	5		
P_{17}		6			P_{46}				17	B_{24}				24	E_6		15		
P_{18}		7			P_{47}				18	B_{25}				29	E_7		16		
P_{19}		11			P_{48}				22	C_1	9				E_8		27	27	
P_{20}		12			P_{49}				23	C_2	14				E_9		10	10	
P_{21}		17			P_{50}				27	C_3	19				E_{10}			15	
P_{22}		18			P_{51}				28	C_4	20				E_{11}			20	
P_{23}		23			B_1	8				C_5	29				E_{12}			21	31
P_{24}		24			B_2	13				C_6	33	35			E_{13}			5	5
P_{25}		28			B_3	18				C_7	4	4			E_{14}				10
P_{26}		29			B_4	25				C_8		14			E_{15}				16
P_{27}		32	33		B_5	28				C_9		21			E_{16}				21
P_{28}		1	2		B_6	32	34			C_{10}		26			E_{17}				26
P_{29}			6		B_7	3	3			C_{11}		31	35						

註：P為物理領域試題，B為生物領域試題，C為化學領域試題，E為地球科學領域試題；數字為各卷題號。

自然科成就之相關。

(五)以積差相關瞭解間隔三週之重測信度。

(六)以試題參數比較法，瞭解試題有無性別的差異試題功能現象。

(七)建立常態轉換的常模分數。

參、研究結果

本研究以試題反應理論進行測驗發展，研究成果乃分兩部分呈現，第一部分呈現題庫試題以水平等化方式實地測試的結果；第二部分則呈現正式測驗的試題品質檢驗與常模建置結果，詳細內容分述如下：

一、題庫品質的分析結果

對於經過內容與文化審查後而保留的試題，進一步蒐集學生實徵表現作為試題品質的參照。研究者將題庫內的試題分為甲、乙、丙、丁四卷進行預試，各卷中皆納入各學科領域不同難度的試題做為共同試題，以水平等化方式取得不受樣本影響的試題參數。

(一)傳統試題分析結果

研究者對四式測驗進行信度，與效標關連效度檢驗，以瞭解題庫試題品質。首先在內部一致Cronbach's α係數與重測信度部分，由表11可知，四卷的內部一致Cronbach's α係數介於 .720到 .783之間；重測信度介於 .810和 .872之間，整體而言，試題有好的內部一致性與時間穩定性。另外，盱衡學術性向測驗與成就測驗在能力光譜上的趨近性，乃檢驗基本學力測驗的自然科成績與四式測驗的相關性，結果關聯效度介於 .698與 .733之間，表示性向測驗與自然科成就表現之間有中高度關連。

(二)試題反應理論分析結果

表12為IRT模式單向度性的檢定結果，可以發現四式測驗四分相關矩陣特徵值，λ_1明顯高於λ_2，而λ_2與λ_3之差距就不明顯，此結果可支持測驗為測量單一主要向度，符合應用IRT模式的要件。

四式測驗彼此有共同試題作連結，在進行試題等化分析後，所有題庫內試題皆有IRT的鑑別度(a參數)、難度(b參數)與猜測概率(c參數)三參數，並繪製題目特徵曲線(Item Characteristic Curve, ICC)與試題訊息曲線(Item Information Curve, IIC)，由表13全部題庫內測驗之三參數IRT試題分析結果，可知a參數的平均值為0.856，b參數的平均值為0.706，c參數的平均值為0.202。

二、正式測驗分析結果

(一)正式卷架構與計分

根據IRT的選題步驟，研究者以最大訊息法(maximum information)在預定目標訊息函數下選擇合適的試題，另方面亦考慮內容

表11：預試階段四卷試題品質檢驗數值表

測驗別 / 卷別	Cronbach's α係數	預試重測信度	效標關聯效度
甲卷	.771	.842	.698
乙卷	.720	.810	.712
丙卷	.783	.872	.717
丁卷	.743	.850	.733

表12：四分相關矩陣特徵值

測驗別	λ_1	λ_2	λ_3
甲卷	6.640	2.037	1.870
乙卷	6.755	2.090	1.680
丙卷	6.435	2.057	1.975
丁卷	6.675	1.900	1.772

表13：三參數IRT試題分析摘要表

試題參數	平均數	標準差	最小值	最大值
a參數	0.856	0.410	0.250	3.487
b參數	0.706	1.192	-1.964	4.393
c參數	0.202	0.049	0.010	0.297

效度，兼顧科目比例分配、測驗能力比例進行選題組卷，由於物理相較於其他學科更能夠鑑定出學生之邏輯思考能力，故物理領域試題較多。表14是正式測驗學科領域暨測量能力雙向細目，由表中可知物理、化學、生物與地球科學的題數分別為：16題、6題、7題、6題。施測時間為50分鐘，計分除了可採用傳統答對題數計分法，每答對1題得1分外；若有需要亦可使用IRT能力估計法，將受試者反應組型輸入電腦，以能力參數估計軟體來估計受試者之能力。

此外，由表15可知選入正式測驗的試題，a參數的平均值為0.809，b參數的平均值為1.096，c參數的平均值為0.190(測驗特徵曲線圖詳如圖3)，由於本測驗的試題皆為五選一的選擇題，必然有某種程度的猜對率，c參數平均值0.190正適切地反應出此一合理的猜測度；再者，由b參數平均看來，全測驗試題偏難，表示測驗對於能力好的受試提供較高的訊息與較低的測驗誤差，符合作為資優甄選測驗的目的。

(二)內部一致性信度與重測信度

以常模樣本檢測之內部一致性Cronbach's α係數為 .813；在間隔三週後，以四個班級學生進行重測，重測信度則為.898。顯見試題取樣誤差與時間取樣誤差皆在合理範圍，測驗具有相當程度的可靠性。

(三)與基本學力測驗自然科成績的相關分析

測驗成績與基本學力測驗之自然科成績的相關為 .782 ($p < .01$, $N = 932$)，基本學力測

表14：正式測驗學科領域暨測量能力雙向細目

卷別	領域	能力(高層思考)		題數
		經驗—歸納推理	假設—演繹推理	
正式測驗	物理	7	9	16
	化學	3	3	6
	生物	3	4	7
	地科	2	4	6
	合計	15	20	35

表15：試題三參數IRT試題分析摘要表

試題參數	平均數	標準差	最小值	最大值
a參數	0.809	0.462	0.373	3.487
b參數	1.096	1.131	-0.482	3.850
c參數	0.190	0.051	0.056	0.207

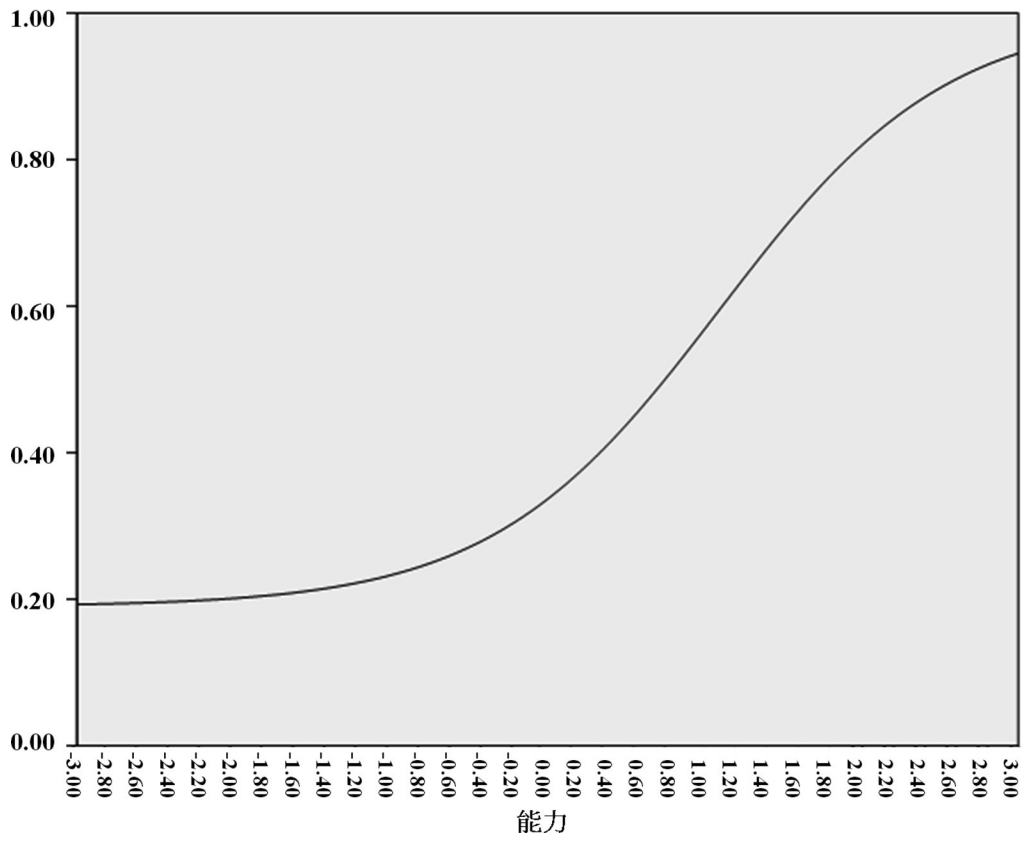

圖3：測驗特徵曲線圖

驗測得學生的自然科能力與本研究之性向測驗測得的自然科學習能力有中高度的相關連性，顯示本研究所編製的性向測驗與自然科學習成就有關，符合學業性向測驗需建構於學生在學科背景知識之期待。

(四)測驗在性別的試題差異試題功能

一般而言科學被視為有利於男生的學科，研究者在命題與審題歷程，已特別注意試題在性別群體的不偏性，以免造成能力估計時的系統性偏差；就IRT的觀點，能力相同的不同群體受試，答對某一試題的概率應該一樣，兩者間若有差異，便是試題DIF，本研究採此一觀點來驗證二式測驗試題無性別DIF，對男、女生受試都是公平的，也以男女生群體所估計試題特徵曲線的差異情形，來探討男女生群體在各試題的得分是否受到不同變異來源的影響。

研究者分別以男生及女生受試樣本估計a, b, c參數後，計算兩群體的b參數相關為 .968，在散佈圖中圖形幾近一直線，兩群體在全測驗似無試題偏失，為進一步瞭解個別試題在兩群體是否有較大的估計誤差和試題偏失，乃對兩群體的試題參數進行連結，整體測驗在男生群體的試題難度平均數為0.888，在女生群體試題難度的平均數為1.216，對女生而言試題難度比男生高了0.328，調整兩群體試題難度參數的平均數相同後，計算兩群體所有試題難度參數之差異與標準誤，結果發現男、女生在I13 (z = 3.22,

$p < .01$)與I25 ($z = 2.06$, $p = .04$)的難度參數差異達 .05顯著水準，表示兩試題似乎有DIF存在，進一步根據ETS (Educational Testing Service)的DIF分類系統，衡量兩試題之Δ_{MH}，I13Δ_{MH}絕對值為0.4，I25Δ_{MH}絕對值為0.6，其DIF現象皆為輕微程度。圖4與圖5是男女群體在I13與I25的試題特徵曲線，兩試題皆呈現在男、女生能力相當的情況下，女生作答成功的機率比較高。

雖然DIF分析結果，性別似為少數試題分數的差異來源，但其DIF現象屬於可以忽略的輕微程度，研究者也進一步重新審視試題，未發現對性別不公平的內涵。所以自然科學性向測驗的試題，在性別間仍具有參數估計不變性，尚不構成性別的偏失(bias)。

(五)常模建置

本測驗除了提供百分等級常模外，亦提供常模轉換後的*T*分數，表16為測驗分數常模對照表。

肆、結論

隨著《資優教育白皮書》在2008年初公布(教育部，2008b)，我國資優教育的推展工作，有了清晰的輪廓與明確的努力方向。而在提供服務之前，需要先知道有需求的對象何在？因此本研究發展適用於高級中學的自然科學術性向測驗，以協助高中篩選出具有科學學習潛能學生，以提供他們適切的安置與資優教育方案，促其潛能的充分發展。在測驗編製上所獲致的具體結論如下：

一、測驗計有三十五題，內容分成生物、物理、化學、地科四個領域。測驗能力向度，則以科學思考與推理能力為核心，分為歸納推理與演繹推理兩部分。所有試題皆以學生熟知的素材或背景知識為前提進行命題，以瞭解學生能否對科學方面知識融會貫通，而進一步進行因果推理或根據既有的科學定律去解釋日常現象。最後，學生需運用兩步驟以上的思考方能解題。

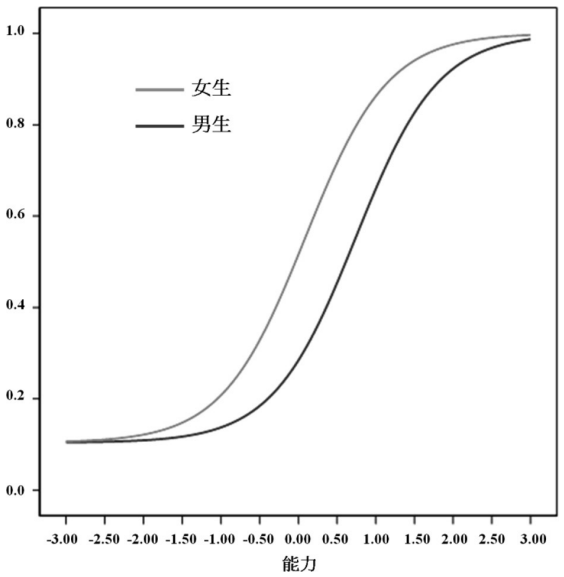

圖4：男女群體在I13的ICC

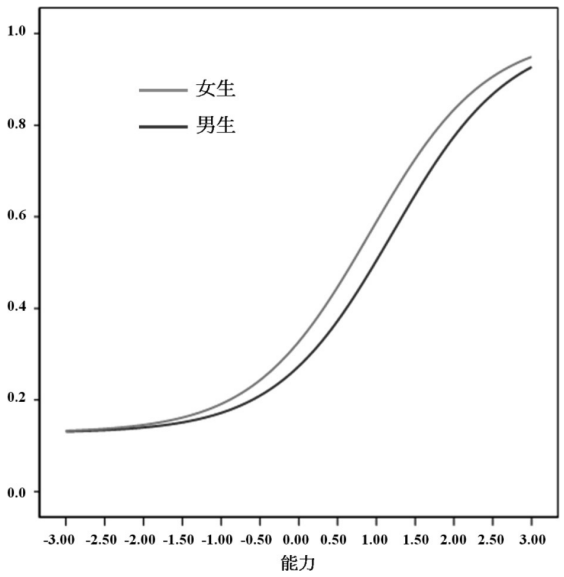

圖5：男女群體在I25的ICC

表16：測驗分數常模對照表

分數	正式測驗		分數	正式測驗	
	百分等級	*T*分數		百分等級	*T*分數
1.00	1以下	19	29.00	97	69
2.00	1以下	24	30.00	98	70
3.00	1	27	31.00	98	72
4.00	2	29	32.00	99	73
5.00	2	30	33.00	99以上	74
6.00	3	32	34.00	99以上	76
7.00	5	33	35.00	99以上	81
8.00	7	35			
9.00	9	36			
10.00	11	38			
11.00	14	39			
12.00	17	40			
13.00	21	42			
14.00	26	43			
15.00	31	45			
16.00	38	47			
17.00	46	49			
18.00	55	51			
19.00	62	53			
20.00	68	55			
21.00	73	56			
22.00	79	58			
23.00	84	60			
24.00	88	62			
25.00	91	63			
26.00	93	65			
27.00	95	66			
28.00	96	68			

二、測驗發展的過程採用題庫建置的流程來進行，考量一個試題充足的優良題庫並非一蹴可幾，本研究完成的測驗，其試題參數的取得，來自全國高一學生群體的代表性樣本，未來可將本測驗的試題作為已校準題目，採用題目預先等化設計方式，逐步透過共同試題，使不同試卷的題目參數，化為相同量尺單位，以擴充題庫規模。

三、本研究將IRT理論用於測驗的發展，在選題組卷時除了考慮學科學習比例，及各題項的認知評量內容之外，也以對能力在+1以上學生能提供最高訊息量為目標來組題，使施測題目數可以減少，且對

高能力者的能力估計值標準誤(standard error)較小。

四、測驗內部一致性信度的Cronbach's α係數為 .813，在間隔三週之重測信度則為 .898。顯見其試題取樣誤差與時間取樣誤差皆在合理範圍，具有相當程度的可靠性。在以基本學力測驗之自然科成績作為效標，其效度也獲得驗證。

五、試題在性別的DIF檢驗結果，發現測驗有兩題呈現達顯著性的試題差異功能現象，女生相對較有優勢，但因屬於輕微DIF，並未有嚴重的試題偏失。

六、本測驗因應教育部(2007)要求資優生鑑定工具應以全國常模作參照，以全國(不含離島)高一學生為母群，取得1,082位學生為測驗常模樣本，建立百分等級、常態*T*分數常模，以作為測驗使用者有效的常模對照。

整體而言，本測驗以嚴謹的過程進行編製，測驗的品質也佳，適合用於篩選有科學性向的學生接受資優教育服務。測驗常模建置時程與目前全國高中資優班學生的甄試時間接近，也提升了使用效度。如果未來能進一步蒐集經本測驗鑑定入班的資優生與一般學生的學業成就表現，將獲得更豐富的效度證據。此外，103年實施十二年國教後，部分學校欲進行特色招生或資優班入學，亦可採用本測驗來瞭解學生的科學學習潛能。最重要的是，本研究的完成，提供未來可以IRT進行不同式測驗題目參數之量尺化工作，俾便逐步擴充題庫。

致謝

本研究為教育部委託之「高級中學自然科學術性向測驗編製」專案，承教育部特殊教育工作小組經費補助，謹此致謝；另外也對林堇梁與林宛萱小姐擔任本研究的主試人員，協助施測工作的辛勞表達謝意。

參考文獻

1. 大學入學考試中心(2001)。**學科能力測驗試題差別功能分析**。臺北市：作者。
2. 大學入學考試中心(2005)。**九十至九十三學年度學科能力測驗、指定科目考試試題差別功能檢核計畫**。臺北市：作者。
3. 毛國楠、盧雪梅(2003)。**中學多元性向測驗**。臺北市：心理。
4. 余民寧(2009)。**試題反應理論(IRT)及其應用**。臺北市：心理。
5. 余民寧、謝進昌(2006)。國中基本學力測驗之DIF的實徵分析：以91年度兩次測驗為例。**教育學刊**，**26**，241-275。
6. 吳裕益(出版中)。**測驗理論**。臺北市：五南。
7. 李秉書(2003)。**高一學生選組及其學業預測之研究——以自然組物理科為例**。未出版之碩士論文，國立臺灣師範大學物理學系，臺北市。
8. 身心障礙及資賦優異學生鑑定辦法(2012年9月28日)。
9. 特殊教育法(2009年11月18日)。

10. 郭靜姿、吳淑敏、黃玉蓮(2006)。**高級中學數學及自然性向測驗指導手冊**。臺北市：教育部特殊教育工作小組。
11. 教育部(2007)。**修正縣市政府資優學生鑑定過程適法性檢核表**。台特教字第0960158634號，2007年10月17日。
12. 教育部(2008a)。**普通高級中學課程綱要總綱**。臺北市：作者。
13. 教育部(2008b)。**資優教育白皮書**。臺北市：作者。
14. 黃堅厚、盧欽銘編(1986)。**高中系列學業性向測驗**。臺北市：中國行為科學社。
15. 路君約、盧欽銘、歐滄和編(1994)。**多因素性向測驗**。臺北市：中國行為科學社。
16. 歐滄和、路君約編(2003)。**多向度性向測驗組合**。臺北市：心理。
17. 盧雪梅、毛國楠(2008)。國中基本學力測驗自然科之性別差異和差別試題功能(DIF)分析。**測驗學刊**，**55**(4)，725-759。
18. 簡茂發、吳鐵雄、吳清基、劉奕權、邱美玉、王俊明等(1994)。**高一性向測驗**。臺北市：中國行為科學社。
19. Linn, R. L., & Miller, M. D. (2009)。**教育測驗與評量**(*Measurement and assessment in teaching*; 9th ed.；王振世、何秀珠、曾文松、彭文松譯)。臺北市：雙葉書廊。(原作出版於2005年)
20. Aberg-Bengtsson, L. (2005). Separating quantitative and analytic dimensions in the Swedish scholastic aptitude test (SweSAT). *Scandinavian fournal of Educational Research*, *49*(4), 359-383.
21. Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.
22. Angoff, W. H. (1993). Perspectives on differential item functioning methodology. In P. W. Holland & H. Wainer (Eds.), *Differential item functioning* (pp. 3-25). Hillsdale, NJ: Lawrence Erlbaum Associates.
23. Council of Ministers of Education. (2005). *Report on Science III Assessment, SAIP 2004*. Toronto, Canada: Author.
24. DeMars, C. (2010). *Item response theory*. New York: Oxford University Press.
25. Dorans, N. J., Pommerich, M., & Holland, P. W. (Eds.). (2007). *Linking and aligning scores and scales*. New York: Springer.
26. Dunbar, K., & Fugelsang, J. (2005). Scientific thinking and reasoning. In K. J. Holyoak & R. Morrison (Eds.), *The Cambridge handbook of thinking & reasoning* (pp. 705-726). New York: Cambridge University Press.
27. Education Quality and Accountability Office. (2008). *Pan-Canadian Assessment Program (2007) - Ontario Report: English-Language Students*. Toronto, Canada: Author.

06-侯雅齡.indd 210 2013/7/12 上午 11:46:34

28. Embretson, S. E., & Reise, S. P. (2000). *Item Response Theory for psychologists*. Mahwah, NJ: Lawrence Erlbaum Associates.
29. Hambleton, R. K., & Swaminathan, H. (1989). *Item response theory: Principles and applications*. Hingham, MA: Kluwer-Nijhoff Pub.
30. Han, L., & Hoover, H. D. (1994). *Gender differences in achievement test scores*. New Orleans, LA: National Council on Measurement in Education. (ERIC Document Reproduction Service No. ED 369 816)
31. Hogan, K., Nastasi, B. K., & Pressley, M. (1999). Discourse patterns and collaborative scientific reasoning in peer and teacher-guided discussions. *Cognition and Instruction*, *17*(4), 379-432.
32. Holland, P. W., & Wainer, H. (Eds.). (1993). *Differential item functioning*. Hilisdale, NJ: Lawrence Erlbaum Associates.
33. Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). New York: Springer.
34. Lawson, A. E. (1994). *Science teaching and the development of thinking*. Belmont, CA: Wadsworth.
35. Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. HilIsdale, NJ: Lawrence Erlbaum Associates.
36. Miller, F. P., Vandome, A. F., & McBrewster, J. (Eds.). (2009). *Item response theory*. Beau Bassin, Mauritius: Alphascript.
37. Millman, J., & Arter, J. A. (1984). Issues in item banking. *Journal of Educational Measurement*, *21*(4), 315-330.
38. Stage, C., & Ögren, G. (2004). *The Swedish Scholastic Assessment Test (SweSAT)*. Umeå, Sweden: Umeå University.
39. Thagard, P. (2005). How to be a successful scientist. In M. E. Gorman, R. D. Tweney, D. C. Gooding, & A. P. Kincannon (Eds.), *Scientific and technological thinking* (pp. 159-171). Mawah, NJ: Lawrence Erlbaum Associates.
40. Vale, C. D. (1986). Linking item parameters onto a common scale. *Applied Psychological Measurement*, *10*(4), 333-344.
41. Vale, C. D. (2006). Computerized item banking. In S. D. Downing & T. M. Haladyna (Eds.), *The handbook of test development* (pp. 261-285). Mahwah, NJ: Lawrence Erlbaum Associates.
42. Zaaiman, H., Flier, H. V. D., & Thijs, G. D. (2000). Selection as contract to teach at the student's level: Experiences from a South African mathematics and science foundation year. *Higher Education*, *40*(1), 1-21.

06-侯雅齡.indd 211 2013/7/12 上午 11:46:34

43. Zhang, L., & Manon, J. (2000). *Gender and achievement - Understanding gender differences and similarities in mathematics assessment*. Paper presented at the annual meeting of the American Educational Research Association. New Orleans, LA.

44. Zimmerman, C. (2005). *The development of scientific reasoning: What psychologists contribute to an understanding of elementary science learning*. Washington, DC: National Academy of Sciences.

科學教育學刊
2013, 第二十一卷第二期, 189-213
DOI: 10.6173/CJSE.2013.2102.03

Chinese Journal of Science Education
2013, 21(2), 189-213

The Development of Natural Science Academic Aptitude Tests for High School Students

Ya-Ling Hou
Department of Special Education, National Pingtung University of Education

Abstract

Scientific talent is an important category of giftedness. An excellent indicator of scientific talent is standardized aptitude tests. The purpose of this study was to develop a Natural Science Aptitude Test for identification of scientific talented students. The test developer established a common scale for the three parameters logistic test model by test equating. Tests editing applied item response theory. The content domains included Physics, Chemistry, Biology and Earth Scienceknowledge. Coefficient of internal consistency (Cronbach's α) was .813. Coefficients of stability for three weeks achieved .898. The criterion-related validity evidence from estimate of the correlation between "The Basic Competence Test" scores and "The Natural Science Aptitude Tests" scores was .782. In the study of gender differential item functioning (DIF), results showed that all of items did not display gender DIF. Percentile rank and normal *T*-score norm of test was based on 1,033 High School Students.

Key words: Natural Science, Test Development, Gifted/Talented Students, Item Response Theory

06-侯雅齡.indd 213 2013/7/12 上午 11:46:34