特殊教育研究學刊
民 112，48 卷 1 期，97-128 頁
DOI: 10.6172/BSE.202303_48(1).0004

科學資優生鑑定之性向測驗題庫發展及運用

侯雅齡
國立屏東大學特殊教育學系
教授

我國在科學學術性向資賦優異學生的鑑定，有標準化科學性向測驗不足，以及測驗題目外洩的風險，影響鑑定結果的效度。本研究目的在建置國中科學性向測驗題庫與高中科學性向測驗題庫，並運用題庫中的試題，組成可提供資優生鑑定用的測驗。科學性向題目採取科學能力、認知歷程與學科知識三向度的構念進行命題，其中科學能力包含科學地解釋現象、評估與設計科學探究、科學地詮釋資料與證據與科學問題解決四種科學學習能力。所有試題皆經過嚴格的內容審查，並以試題反應理論（IRT）透過水平等化設計發展題庫，建立各試題的鑑別度、難度、猜測度三參數。國中科學性向測驗題庫，先透過全國 4,663 位學生，建立 140 題的校準題庫，再從中編組一套 7 題的定錨測驗做為後續擴增試題等化之用，使不同測驗施測不同的樣本，也能與校準題庫有相同基準的參數。後續再以每式測驗 34 題組卷，總計擴增了 309 題，IRT 參數顯示此題庫的試題有良好鑑別度、難度值中間偏難，及合理猜測度。高中科學性向測驗題庫，先透過全國 3,702 位學生，建立 274 題的校準題庫，再從中編組一套 10 題的定錨測驗做為後續擴增試題等化之用。後續以每式測驗 50 題進行組卷，總計擴增了 412 題，IRT 參數顯示此測驗具有良好鑑別度、難度值中間偏難，及合理猜測度。以資優學生之學校自然科學期成績，作為效標關聯效度考驗，結果符合期待。資優生與一般生在測驗各向度的表現，皆有顯著差異，顯示區辨效度良好。最後也考量實際鑑定需求，由題庫中選 50 題組成測驗，並建立九年級學生的百分等級及常態化的 T 分數常模，以說明題庫的運用方式。

關鍵詞：科學性向測驗、測驗編製、資優學生、題目反應理論、題庫發展

* 本文作者通訊方式（yalingho@mail.nptu.edu.tw）。

** 致謝：本文為教育部國民教育及學前署委辦計畫之部分成果，感謝洪振方教授協助科學性向理論架構、所有參與命題與審題的委員，以及協助到校的多位研究助理。

緒論

從近十年來《特殊教育統計年報》（教育部）的資料來看，六大類資賦優異學生人數，在國、高中階段，以學術性向資賦優異類最多，其中，國中學術資優生占整體國中資優生比例介於 42.5% 至 51.6%，高中學術資優生占整體高中資優生比例為 45.6% 至 48.3%。而學術性向包含的領域有語文、數學、社會科學及自然科學，目前各縣市所成立的國中學術性向資優（資源）班又以數理資優生人數最多，幾乎占整體國中學術資優生的一半（50.7%），近十年來人數從 2722 人增加到 4700 人；至於全國各高中的學術性向資優班也以數理資優班居多，大約占整體高中學術資優生的三分之一（介於 31.8% 至 34.4% 之間），因此對可用於鑑定國中與高中的數理資優學生的測驗工具，需求特別殷切。

根據身心障礙及資賦優異學生鑑定辦法（2013，教育部）第二條第二款規範「資賦優異學生之鑑定，應以標準化評量工具，採多元及多階段評量……」。在此條文中，強調評量工具需為標準化測驗，亦即測驗需按照標準化程序而編成：亦即在題目的選擇，需經過試題的分析，至於測驗的實施、計分和解釋，有一定的程序，且需有常模提供參照，評量程序則需為多元、多階段評量。

張正芬（2014）在資優鑑定辦法的說明手冊中，有關學術性向資賦優異學生鑑定流程，制定初審、第一階段評量（初選）、第二階段評量（複選）與綜合研判等階段，其中初選與複選兩個階段，建議宜分別採用性向測驗及成就測驗／實作評量等，以評估學生未來學習潛力與過往表現，避免僅採用單一成就測驗來鑑定學生，使具有資優潛能的學生皆能受到適性的教育。由於通過初審的學生一般都已展現相對優異的學習成就表現，但是對於學生未來學習潛力的評估就需要使用性向測驗，學術性向測驗不同於多元性向測驗，偏屬於特殊性向測驗的範疇，目的在了解受試在特定學術領域的特殊潛能，我們相信潛能愈高者，在該學術領域相關活動的學習及表現成功的可能性也愈高。因此，用於科學資優生鑑定的性向測驗，目的在協助篩選有科學性向的學生，接受資優教育的服務以開展其潛能，使其在未來的學科學習有較佳的表現，故測驗的內容需與學生未來學習學科的內容有所關聯（Zaaiman et al., 2000）並從中考量科學能力（侯雅齡，2010，2013）。

2009 年 11 月修正公布之《特殊教育法》（教育部），新增列了高級中等教育階段特殊教育的提供，對於資賦優異學生應給予充實學習之機會。雖然，教育部因應十二年國民基本教育實施，自 96 學年度起實施高中職優質化、均質化等政策，協助學校建立優質特色，希望打破以往外界所注重明星學校或傳統名校的迷思，讓眾星拱月（校本群聚菁英）的明星高中思維，轉為群星爭輝（多元群聚菁英）為目標（教育部國民及學前教育署，2017），多年來頗有成效，許多國中學生皆能適性就近選擇高中就讀。但是，不可否認前端高中能力分校的狀況仍然存在，目前設有「數理資優班」的學校也以傳統的明星高中居多，再者，援《特殊教育法》高中資優學生的安置型態採取集中編班的方式，因此，競爭特別激烈。根據國民及學前教育署特殊教育評量工具借用銷售及管理要點，用於資優鑑定的測驗，屬於保密測驗並不會公開題目，但卻常屢遭有心人士刻意蒐集測驗內容，舉例來說，2017 年 8 月就曾發生過

全國 41 所高中數理資優班聯合入學考試的洩題事件，讓考試的公平性與測驗的效度，皆受到極大的考驗（自由時報，2017）。

另外，在國中部分，每年全國各縣市參與數理資優鑑定的學生眾多，且辦理鑑定的時間不一，以致於測驗的曝光度高，保密性普遍不佳，因此測驗能重複使用的次數極為有限，堪用時間也短。

事實上，為符合測驗標準化的要求，每一份性向測驗的編製耗時又所費不貲。但是在僧多粥少、家長望子成龍望女成鳳的現實下，資優學生的鑑定的確存在著許多挑戰，在大眾尚未能對資優教育服務有正確的態度前，對於此類高風險的測驗應改變目前編製固定題目的測驗方式，未來應建置科學性向題庫，透過足夠大的題庫來降低試題外洩的風險，每一年需要測驗的時候，由題庫選題組卷，使每一次測驗題目皆不相同，以杜絕坊間補習班透過蒐集考生應試後對題目記憶，嘗試還原測驗題目的企圖（侯雅齡，2013；Crisp et al., 2020；Vale, 2006）。

研究者以為，從發展資優鑑定工具的需求性、資優鑑定流程中的適切性、到資優鑑定過程必須保持的公平嚴謹性，未來若是仍照過去的做法，只是輪用目前已有的性向測驗，洩題的風險以及鑑定效度的疑慮依舊無法消弭，為了維護測驗的有效性，確保鑑定效度讓資優學生能接受適性的教育，題庫建置勢在必行。若題庫可以透過後續增修的形式來維護，不僅可以節省不斷研發新測驗的成本，也避免過往因為安全考量，刻意避免測驗中雷同或重複的試題，而造成測驗難度不斷的提高，影響測驗的品質（侯雅齡，2013；Crisp et al., 2020），因此，本研究目的乃在建構一個可以持續發展的國中科學性向測驗題庫與高中科學性向測驗題庫，此題庫未來可提供科學資優學生鑑定之用。

一、題庫建置的方式與挑戰

題庫建置是一種測驗發展的策略，前期透過編寫大量獨立的題目並彙整，成為一個題目池（item pool）或題目庫（item bank），接者，再將試題池的題目經過專業的審查、修訂，最好能透過預試，取得實徵的數據後，以古典測驗理論（classical test theory, CTT）或試題反應理論（item response theory, IRT）來取得題目的難度（b 參數）、鑑別度（a 參數）、以及猜對率（c 參數）等訊息並做題目校準 (calibration），確保題庫內題目的品質（Crisp et al., 2020）。為了確保題庫有重複使用的成本效益並延長題庫使用期限，需要定期增、刪題目，且最好能有校準與等化的機制。

雖然 CTT 的理論內容較容易了解，且為多數測驗發展使用的基礎理論，但是以 CTT 來發展測驗，所獲得的難度與鑑別度數值會隨樣本特性改變，相同的題目給能力高低不同的群體施測後，會獲得不同的題目參數。CTT 的等化方法有線性等化法（linear equating）與百分位數等化法（equipercentile equating），前者會假設測驗之間的難度相近、偏態與峰度不變的狀況下來進行等化，而後者則是只能在測驗與測驗的百分位數相同時，直接經由對照百分位數來等化。

如果想獲得不受實施群體能力影響的獨立題目參數，則必須採用 IRT 來建構測驗、透過等化的方式，讓題目建立在相同的量尺上，以利相互比較與使用（侯雅齡，2010，2013）。以下乃先討論如何採用 IRT 來建構題庫。

（一）考量題庫使用目的與規模

在題目發展之初，應先依題庫研發目的，擬定題目測量能力的雙向細目表、規範命題的範圍、設定題目的條件，再邀請命題委員開發題目。好的題庫必須依研發目的建置足量的題目，往往須滿足的條件越多，題庫所需的規模越大。一般而言，在題目需求數量上，過往研究建議每個概念至少要包含十個題目，若有區分單元，每一單元內容至少要包含五十題，最少有一百到二百題難度均勻分布且具有合理鑑別度的試題，且如果未來組卷的要求越多，例如，一份測驗須包含多少題目、要由幾個領域組成、題目難度的分配比例等，題庫內的題目數量就需要越多（余民寧，2009，2011）。但也有研究指出，即使題庫內的題數達到一份測驗所需題數的二十倍，仍然可能難以滿足某些測驗在組卷的要求（Crisp et al., 2020），對此，Bramley 等人（2019）建議，開始啟動題庫題目研發時，最好先依照未來實際組卷的測驗題型與領域比例去研發題目，這樣可以降低題庫所需的題目數，快速滿足組卷的需求。

（二）應用等化方式擴增題庫規模

等化方式有水平等化（horizontal equating）與垂直等化（vertical equating）兩種方式：水平等化是利用兩個或兩個以上的測驗去測量同一特質時，測驗間具有相近的能力分配與試題難度，使測驗間的關係如同複本。例如多益（TOEIC）這種可以在一年內舉辦多次測驗的考試，經由等化程序後，接受不同份測驗的不同受試者可以互相比較成績高低。至於垂直等化則是利用兩個或兩個以上的測驗去測量同一特質時，測驗間有不同的能力分配與試題難度，可用來檢測某特質在不同年齡、年級的分配情形，常見於成就測驗中劃分不同難度及適用的年齡或年級。

建置題庫時需要大量的題目，我們不可能讓相同的受試接受大量的題目，受試者生理疲勞與心理抗拒的因素都會造成偏誤的題目訊息。只要測驗具有相同的結構、測驗間的受試者沒有差異，可提供不同的受試群體適量題目的測驗，透過水平等化的方式，將兩份或兩份以上的不同測驗得分轉換到同個量尺（Dorans & Holland, 2000；Hambleton & Swaminathan, 1985；Lord, 1980）。

IRT 使用的等化方法有迴歸法（regression method）、平均數與標準差法（mean and sigma method）、強韌平均數標準差法（robust mean and sigma method）、參數比較法與特徵曲線法（characteristic curve method），會考慮測驗的能力與難度進行運算。其中特徵曲線法等化效果較佳，也較為被廣泛接受，原因在於其作法涉及測驗的鑑別度（discrimination）、難度（difficulty）及猜測度（guessing）三項指標參數，獲得的訊息量豐富且與所測量的群體相互獨立，除了可以較精確地估計不同能力受試者的表現外，更不會因為受試群體不同而改變參數數值，獨立存貯的題目訊息，也讓後續組卷時，可以依據不同目的，來選擇適切的題目組卷（余民寧，2009；侯雅齡，2010，2013；郭伯臣等人，2003）。研究者將常見的等化設計整理於如表 1（余民寧，2009；侯雅齡，2013；Dorans et al., 2007；Kolen et al., 2004），綜合表 1 中各種等化分析特點，考量科學性向的每一試題在作答時都需要較長的思考時間，每一卷題目不宜太多，為了不斷擴充試題需要銜接多分測驗，乃參採侯雅齡（2013）的做法，以定錨測驗設計，在測驗間納入共同試題進行水平等化。

（三）持續修改與增刪更新題庫題目

表 1　常見等化設計

	分組	欲銜接之測驗式數	每組施測的測驗	特點
單組設計	1	2	2	有疲勞及重複練習的影響，干擾第二式的參數結果
對抗平衡隨機組設計	2	2	2	一組先接受 A 式再接受 B 式，另一組則相反，降低前項影響
等組設計	2	2	1	簡化前兩項設計，只分別測驗其中一式測驗，需要更大的樣本數
定錨 (共同) 測驗隨機組設計	2↑	2↑	1+ 定錨	除了在不同群體間施測不同測驗之外，另有一份共同測驗用以同時測量兩組受試者
定錨測驗不等組設計	2↑	2↑	1+ 定錨	相較於前項，此設計組別為不等組或自然產生的分組，有可能出現系統性差異
分測驗預先銜接等化設計	2↑	數個段落組成	1(相同段落 + 不同段落)	將某測驗在成為完整測驗之前，就先等化到另一測驗，再拼接組卷
試題預先等化設計	1↑	1↑	1↑	根據已校準的試題建立新測驗並加以等化，用來擴充題庫

題庫內的題目應不斷的擴充，若有不符合現今所需，也應該持續進行後續的修改或增刪，以延長使用期限（余民寧，2011）。首先，應持續在題庫設定的目標之下，邀請命題委員依題庫目標與命題原則研發題目，並經專家全面的審查與修改，確認符合雙向細目表設計，也注意題目編寫的形式及風格（Crisp et al., 2020）。最理想的狀況，是每個題目皆取得實徵資料，透過等化程序將題目進行校準，建立新題目的題目參數，以不斷擴充題庫規模。再者，題庫內的題目必須不斷地檢視是否進行修改與淘汰，檢視題庫依據的理論構念或法規基礎是否有修正，避免測驗重點或主題無法如實呈現，或出現不合時宜的題目，以及對於已施測過的題目，可以蒐集作答狀況，如答對率、鑑別度等，必要時重新修訂題目後，再考量是否重新納入題庫。

整體而言，透過建立題庫來發展測驗的優勢包含：1. 若是可以重複使用題目，將節省發展測驗的金錢與時間（Malak, 2013；Wood, 1968）；2. 節省組成測驗的時間，甚至可以依照題目難度設定及學生程度，快速組成適當的試卷（Byrne, 1976； Wright & Bell, 1984）；3. 若有依照主題分類題目，還可以進一步依每次測驗、受試者，甚至需求不同，組建不同的測驗來反映不同的資訊（Wood, 1968；Wright & Bell, 1984）。另一個較少被提及的優點是，發展單份測驗相較於編寫題庫，單份測驗更需要時時注意每題目與題目之間的匹配程度、題目與整份測驗的關聯性，可能導致命題者需要放棄一些不錯的題目。而在編寫題庫的題目時，命題者只要專注於發展高品質的題目即可，整份測驗的構念與配置，可於組卷測驗時再做考量（Crisp et al., 2020）。

二、科學性向測驗的內涵與構念

在六類資賦優異中，有關學術領域資賦優異學生的鑑定，主要使用學術性向測驗或成就測驗來做判斷，一般認為，相較於測量學生於學校習得的知識的成就測驗，性向測驗較不受學校學習經驗限制，比較能預測個人未來可能發展的潛在能力（potentiality）。但是 Miller 等人（2013）認為這樣區分過度簡化兩者的關係，在其專書中（Measurement and Assessment in Teaching）從兩種測驗所測量的學習類型以及最有效的預測類型，做深入的探討。就測未來表現來看，Cronbach 的能力測驗光譜將能力測驗分成 A 到 E 五個連續性水準，水準 A 到 B 是成就測驗，水準 C 到 E 是性向測驗。本研究的科學學術性向測驗在測量學生運用科學學識來進行問題解決，並以此預測未來學習科學的成效，故在分類水準上是接近成就測驗性向測驗（水準 C），有別於光譜中水準 E 的普通性向測驗，為測量語文、空間知覺、思考或邏輯推理等內容，類似一般智力測驗；也就是科學性向測驗偏向測量學生在自然領域的特殊潛能，尤其，資優學生鑑定的目的在篩選有需求的學生，希望這些學生在接受資優教育後，其未來的學科相關領域的學習表現更為傑出。因此，科學性向測驗的內涵必須與學生未來學習學科的內容相關（李秉書，2003；Zaaiman et al., 2000）。

侯雅齡（2010）編製的國民中學科學性向測驗，採用了學科領域與高層思考能力兩向度，高層思考能力包含：1. 科學重要概念理解能力，2. 圖表資料解讀與運用能力，以及 3. 推理能力、分析比較能力與問題解決能力。

侯雅齡（2013）編製的高級中學科學性向測驗，也是採用兩向度構念，分別是學科領域與高層思考與推理能力的測量，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，則以「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。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，綜合出新的事物或資訊。

不管是國民中學科學性向測驗或是高級中學科學性向測驗，共通之處在強調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。希望了解資優生如何由已有資料（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。

由經濟合作暨發展組織（Organization for Economic Co-operation and Development，簡稱 OECD）所委託的國際學生能力評量計　畫（Programme for International Student Assessment，簡稱 PISA）運用素養（literacy）的觀點來設計測驗，主要目標之一，在評估學生面對變動快速社會的能力，該評量以 15 歲學生為對象，在科學能力的測量包含能科學地解釋現象、評量與設計科學探究、解讀科學數據及舉證科學證據的能力。由測量內容來看對於瞭解其學術潛力及學習態度，有相當的解釋力。

再因 Bloom 認知歷程層級分類（Anderson & Krathwohl, 2001；Bloom et al., 1956）來看，下層屬於基礎思考，越往上越傾向高層次的認知能力，學生要先理解才能進一步應用與分析，其中理解能力顯示對概念含義的理解，評量時往往需要學生進行描

述、解釋等；應用能力指將所學到的概念運用於解決問題，評量時往往需要學生進行演示或解決等；分析能力是解釋概念的組成部分，區分事實和假設，評量時往往需要學生進行推斷、比較；評鑑能力能獨立判斷所學概念的價值、有用性或優勢，評量時往往需要學生進行評估、解釋。

綜合上述，本研究的國中及高中科學性向測驗，分為學科知識、科學能力與認知能力向度，其中在科學能力上，再加上科學問題解決，作為科學性向測驗的理論架構（侯雅齡等人，2018）。

研究方法

一、試題編製構念

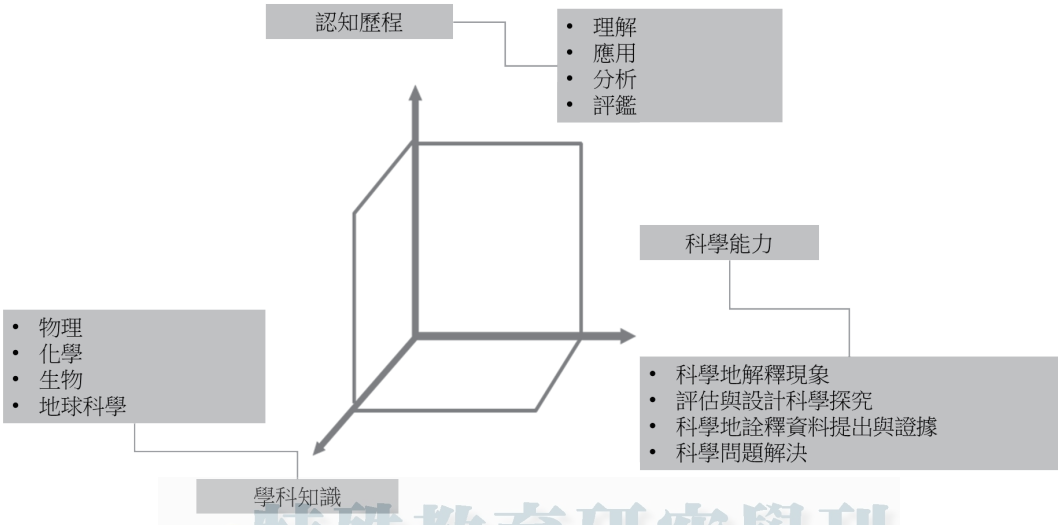
本研究目的在建置科學性向測驗題庫，所有的題目皆以圖 1 的科學性向理論架構作為試題編製的依據，此一架構分為三個向度：學科知識（content knowledge）、認知歷程（cognitive processes）及科學能力（science competences）。

（一）學科知識是以國中、高中自然領域課程的內涵做考量，區分為物理、化學、生物和地球科學四個科目，在研究執行的後期，十二年國教的課程綱要公布並實施，為了呼應十二年國教的理念，在最後一輪命題時，也在四個學科之外增加跨學科的試題。

（二）認知歷程考量性向測驗之目的在了解學生的學習潛能下，以高層次思考能力為主，包含理解（understand）、應用（apply）、分析（analyze）和評鑑（evaluate）四個認知能力。

（三）科學能力則考量科學學習的能力，提出科學地解釋現象（explain phenomena scientifically）、評估與設計科學探究（evaluate and design scientific enquiry）、科學地詮釋資料與證據（interpret data and evidence scientifically）與科學問題解決（scientific problem solving）四種科學能力。

圖 1　科學性向理論架構



二、題庫建置程序

本研究歷時六年，第一年主要工作在確認科學性向試題編製構念，隨後則如圖 2 題庫建製流程，循環不斷地進行試題擴增、試題品質檢核、試題訊息建置的工作，各階段的執行重點茲說明如下：

（一）試題擴增階段

1. 建立試題命題人才資料庫

研究初始乃先建立試題命題人才資料庫，人才來源包含（1）自然科領域國教輔導團中央團及地方團成員，（2）經學者推薦的科教碩士及博士，（3）各縣市推薦之有豐富自然實作經驗的老師，如科展指導教師或社團指導教師，（4）國家博物館以及國家公園的研究人員。

2. 邀請委員進行試題編寫

研究者依學科領域從人才庫中邀請專業人員協助命題工作，每次命題前召開試題編寫會議，提供試題編寫手冊並與命題委員充分討論試題編寫原則，協助命題者能以科學性向理論架構進行命題。

（二）試題品質檢核階段

為確保試題品質，所有收回的試題皆須交由學科專家、科教專家以及測驗專家進行題目審查與修改，並在專家修審後將試題組成試卷，選取小樣本的目標對象進行試做，以進一步回饋文句與題意的適切性。

1. 形式審查

確認命題委員提供之試題是否符合形式要件及命題原則，若有不完整需補正之處（例如，試題圖示不清、題幹未附圖表但卻於題幹敘述提及圖表等），則聯繫命題委員進行修正。

2. 學科專業內容審查

試題編寫分為物理、化學、生物以及地球科學四科目，因此執行學科專業內容審查時也依科目分成四組，各組邀請二至三位具備學科專業知識的專家以及一位同時具備專業學科知識及科學教育知識背景的教授擔任審查委員，共同針對試題之題幹、選項、圖表、答案進行審查，若有題意不清、答案不正確、有爭議或非唯一解，則逕行修改；此外，若試題的選項品質與誘答力不佳也會進行修改；審查時也須確認各試題在認知歷程能力與科學能力之對應。

3. 測驗專業審查

經學科專家審過的試題，全部皆交付測驗專家針對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；也需留意試題是否容易受到某些受試者的特殊反應心向影響，必要時進行調整。

4. 組卷試做

本研究與在地兩所國小與一所國中長期合作，將專家審查通過的試題，組成短卷（約 20 題）由與目標群體相符的學生（每卷約 50 名學生）試做，希望透過實徵的方式，了解學生的觀點與次文化，檢視試題語句敘述方式、題意、遣詞用字等，學生是否能理解，以求試題品質更趨完善。

（三）試題訊息建置階段

1. 建立題目校準題庫

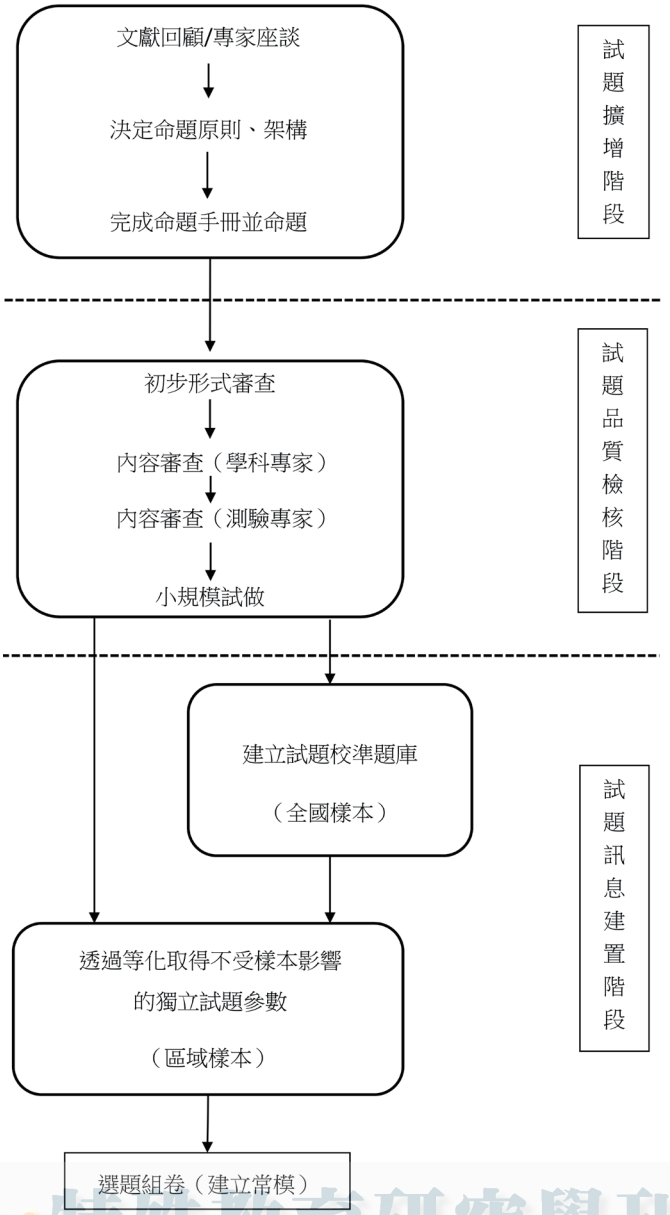
由於本題庫中的題目將做為資優學生鑑定之用，依據鑑定辦法需要以全國學生的施測結果進行參照，因此本研究首先以全國學生為樣本，來建立一組可作為未來試題校準依據的題庫。未來研究只要利用校準題庫內的試題，透過共同測驗使試題參數化為相同量尺單位，即可取得不受樣本影響的獨立試題參數，不斷擴充題庫規模。

2. 透過等化取得不受樣本影響的獨立試題參數

將由新試題組成的測驗透過共同測驗與校準題庫內的試題做連接，即使不同測驗提供給不同的樣本施測，也能讓試題透過水平等化建立在共同的量尺之上，以相互比較與使用，也藉此讓科學性向題庫規模可以斷擴充。所有的試題進行量化分析後，取得 IRT 試題的難度（b）、鑑別度（a）、猜對概率（c）三參數，以及 CTT 的選項答對率等訊息。

題庫內試題的規模只要夠大，未來只要考量施測目的，自題庫中選取具合適訊息函數的試題來組成所要的測驗。

圖 2　題庫建製流程圖



三、等化研究設計

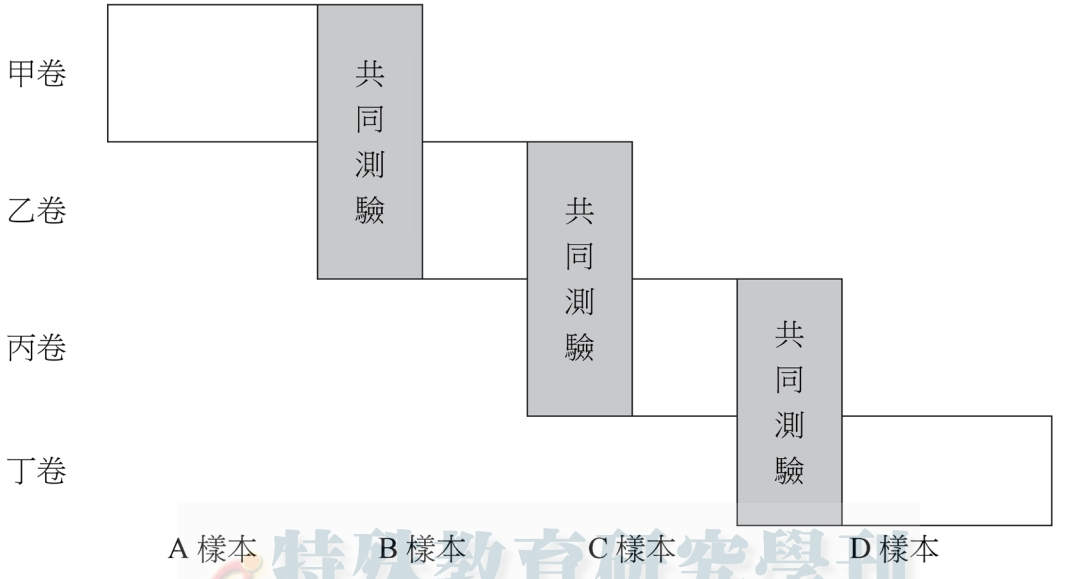
本研究透過跨卷之間的連結試題，以水平等化的方式先建立全國取樣的校準題庫；再由題庫中選取已校準試題組成定錨測驗，做為後續擴增題庫時等化設計，茲將兩種等化設計方式說明如下：

（一）共同測驗隨機組設計

將大量經過質性試題品質檢核過的試題，以量化分析取得測驗的 a、b、c 三參數，採用共同測驗隨機組設計，讓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，以提供不同測驗共同的參照點，減少來自於兩組受試能力差異所造成的等化偏失（equating bias）。

考量科學性向測驗包含物理、化學、生物、地球科學四個學科領域，因此共同測驗規劃有 8 題連結試題，每一學科各有 2 題。三份共同測驗題目分別來自欲連結的甲乙卷、乙丙卷以及丙丁卷，且四卷測驗各自以 A、B、C、D 四群不同樣本施測（詳如圖 3）。

圖 3　共同測驗隨機組設計

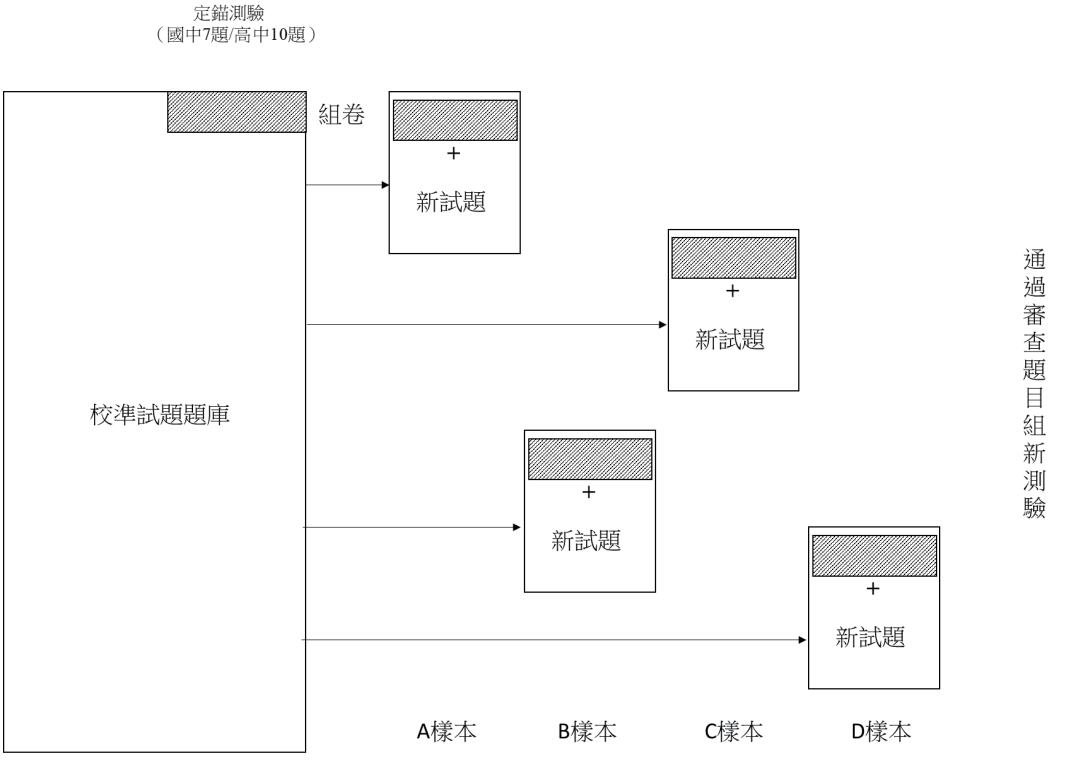


對象為將升國一及國中一年級學生，高中科學性向測驗的對象為升高一的學生，兩族群的生理條件有差異，因此在測驗題數的規劃上不同，國中科學性向測驗為 34 題，高中科學性向測驗為 50 題，在試題分布上，也考量科學性向理論架構去平衡學科比例、認知能力與科學能力。

（二）試題預先等化設計

由圖 4 可知，本研究先以全國學生樣本建立試題校準題庫並取得 IRT 的三參數，後續只要將通過審查的試題組成新測驗，並於新測驗中置入來自校準題庫中的定錨測驗進行水平等化，就能使能力估計值等值，且試題參數的估計值也在同一量尺，逐步擴增題庫試題。考量本題庫應用於資優生鑑定屬於高敏感測驗，應盡可能減低試題編製過程的曝光度，因此乃由校準題庫中選出一組定錨測驗，作為後續新測驗與已校準題庫的連結。考量國中生與高中生生理條件差異，研究者預計未來國中科學性向測驗每卷有 34 題，高中科學性向測驗每卷有 50 題，因此

圖 4　擴充題庫的預先等化設計



二定錨測驗在國中科學性向測驗中包含七題（物理、化學、生物各兩題，地球科學一題），在高中科學性向測驗中則有十題（物理、生物各三題，化學、地球科學各兩題），這些題目除了包含四個學科，也完整包含四個認知能力與四種科學能力。這組測驗只做為等化試題之用，不納入未來組成正式測驗的題庫中。

四、研究參與者

本研究包含建立校準試題題庫的全國樣本、擴增題庫的區域樣本、建立效標關聯效度與區辨效度的樣本，以及組卷後的全國常模樣本。

考量本題庫做為資優學生篩選的目的，施測過程必須標準化也必須嚴格的保護測驗的安全，避免試題外洩，因此，所有的施測工作皆由二至三位經過研究者充分訓練的研究助理擔任主試，在測驗過程中，須詳細記錄施測狀況，若主試人員註記學生於 20 分鐘內即作答完畢與不認真作答者，視為無效樣本；此外，若回收的答案卷經檢核有明顯的反應心向也視為無效樣本。茲將樣本的分述如下：

（一）全國樣本

1. 國中科學性向測驗全國樣本

國中的測驗合計有七卷，分三次進行全國樣本的施測，樣本採分層隨機叢集抽樣，將全國分區以及都市化程度考量後，參酌目標學生分布比例以學校為單位做抽樣，詳細樣本分布狀況如表 2。甲、乙、丙、丁四卷先於 105 年以全國 13 所國小，每所國小各

表 2　國中科學性向測驗全國樣本配置

卷別	國小全國 (I)			國中（入學前）				小計（有效比例）
	學校區域	北區	中區	南區	北區	中區	南區	東區	
	學校數（學校規模大 / 中 / 小）	5 (4/1/0)	4 (3/1/0)	4 (4/0/0)	8 (2/6/0)	8 (2/6/0)	8 (3/5/0)	1 (0/1/0)	
甲卷		133	99	103					
乙卷		133	90	101					
丙卷		132	96	104					
丁卷		130	86	100					
戊卷					356	359	397	21	1,133(99%)
己卷					354	348	395	20	1,117(99%)
庚卷					337	356	392	21	1,106(99%)

表 2　國中科學性向測驗全國樣本配置（續）

卷別	國中（入學後）					合計（有效比例）
	學校區域	北區	中區	南區	東區	小計（有效比例）	
	學校數（學校規模大 / 中 / 小）	12(2/9/1)	9(3/6/0)	8(2/5/1)	1(1/0/0)		68(26/40/2)
甲卷							335 (99%)
乙卷							324 (97%)
丙卷							332 (96%)
丁卷							316 (94%)
戊卷		433	292	343	54	1,122(99%)	2,255(99%)
己卷		422	348	314	57	1,141(99%)	2,258 (99%)
庚卷		393	345	337	60	1,135(99%)	2,241 (99%)

抽取六年級四個班級學生，分別施測甲、乙、丙、丁四卷。各卷樣本數為 316 至 335 人，有效樣本比為 94% 至 99%。整體有效比例 96%。

戊、己、庚三卷於 106 年 4 至 5 月間依臺灣北中南東各區學生數比例，選取 25 所國小六年級學生施測，各卷有效樣本分別為 1,133 人、1,117 人及 1,106 人，有效比例皆為 99%。因應國中資優學生鑑定期程有入學前鑑定與入學後鑑定，因此，除了以六年級學生為對象施測外，考量入學後受試者之母群體為七年級學生，於同年 10 月至 11 月間選取 30 所國中七年級生施測，各卷有效樣本分別為 1,122 人、1,141 人及 1,135 人，有

效比例皆為 99%。

合計國中科學性向測驗校準試題題庫，在入學前的有效全國樣本數有 4,663 人，在入學後的有效全國樣本數有 3,398 人。

2. 高中科學性向測驗全國樣本

高中的測驗合計有七卷，分三次進行全國樣本的施測，樣本採分層隨機叢集抽樣，將全國分區以及都市化程度考量後，參酌目標學生分布比例以學校為單位做抽樣，詳細樣本分布狀況如表 3。首先甲、乙、丙、丁四卷先於 105 年 05 月以全國 17 所國中的九年級學生進行施測，合計取得有效樣本 1,573 人，由於本次於國中會考後施測，有發現少部分受試者配合度低明顯不專心作答，丙卷的有效樣本比例為 85%，略低於其他三卷。於是於 8 月全國高中（職）開學第一週，再分北中南三區，各區選取二所高中 / 職學生為樣本進行再次施測，共計增加 895 人，有效樣本比為 95%。

戊、己、庚三卷，納入東部學校，共計取得 21 所國中 1,234 位九年級學生參與施測，各卷有效樣本數為 415 人、421 人及 398 人，有效樣本比為 92% 至 95%，整體有效比例 93%。

合計高中科學性向測驗校準試題題庫的有效全國樣本數有 3,702 人。

（二）擴增題庫的區域樣本

1. 國中科學性向測驗區域樣本

由於已建有全國樣本的校準試題題庫，因此後續測驗只需以區域樣本進行施測，再透過定錨試題與先前全國資料進行等化，不僅降低施測成本也減少題目曝光度。本研究歷經六年，經前二年的磨合之後，已掌握最佳的施測節奏，利用寒、暑假與學校行政單位聯繫，上學期於 10 至 11 月施測，下學期於 3 至 4 月施測，研究期間於 106 至 109 學年間，共進行五次國中科學性向測驗區域性的測驗，如表 4 所示。

第 1 次施測 A、B、C、D 四卷測驗，以高雄市和屏東縣 14 所國小共 1,361 人為有效樣本，樣本有效比例為 94%，第 2 次施測 E、F、G 三卷測驗，以屏東縣、高雄市和臺南市 13 所國小共 1,097 人為有效樣本，樣本有效比例為 91%，第 3 次施測 H、I、J 三卷測驗，以高雄市、臺南市和嘉義縣市 14 所國小共 1,455 人為有效樣本，樣本有效比例為 95%， 第 4 次施測 K、L、M 三卷測驗，以屏東縣、高雄市 10 所國小共 1,390 人為有效樣本，樣本有效比例為 99%，第 5 次施測 N、O、P、Q 四卷測驗，以臺中市和彰化縣 13 所國小共 1,170 人為有效樣本，樣本有效比例為 96%。總計有 6,473 人參與測驗（樣本有效比例 96%）。

2. 高中科學性向測驗區域樣本

於 107 學年至 109 學年間，每學年進行一次高中科學性向測驗區域性的施測，前兩次施測三卷測驗，第三次施測四卷測驗，如表 5 所示，

第 1 次施測 A、B、C 三卷測驗，以屏東縣、高雄市、臺南市和嘉義縣市九年級學生為對象，共 1,127 人為有效樣本，樣本有效比例為 92%，第 2 次施測 D、E、F 三卷測驗，以屏東縣、高雄市和臺南市九年級學生為對象，共 870 人為有效樣本，樣本有效比例為 95%，第 3 次施測 G、H、I、J 四卷測驗，以屏東縣、高雄市和臺南市九年級學生為對象，共 1,572 人為有效樣本，樣本有效比例為 98%。期間總計有 3,569 人參與測驗，整體有效樣本比例為 95%。

（三）效標關聯效度研究樣本

本研究選取三個八年級資優資源班，共

表 3　高中科學性向測驗全國樣本配置

卷別	國小全國 (I)				國中（入學前）			
	學校區域	北區	中區	南區		北區	中區	南區	
	學校數（學校規模大 / 中 / 小）	5(3/2/0)	6(2/4/0)	6(2/4/0)	小計（有效比例）	2(2/0/0)	2(2/0/0)	2(1/1/0)	小計（有效比例）
甲卷		119	145	141	405(91%)	77	87	72	236(96%)
乙卷		137	122	130	389(90%)	73	78	56	207(92%)
丙卷		132	130	112	374(85%)	79	78	68	225(96%)
丁卷		135	134	136	405(89%)	77	79	71	227(97%)
戊卷									
己卷									
庚卷									

表 3　高中科學性向測驗全國樣本配置（續）

卷別	國中（入學後）				合計（有效比例）
學校區域	北區	中區	南區	東區	
學校數（學校規模大 / 中 / 小）	10(3/6/1)	4(2/2/0)	6(2/3/1)	1(0/0/1)	44(20/22/2)
甲卷					641(94%)
乙卷					596(91%)
丙卷					599(91%)
丁卷					632(93%)
戊卷	175	70	146	24	415(95%)
己卷	162	81	156	22	421(92%)
庚卷	148	109	141	0	398(93%)

76 位資優學生為樣本，以其自然科領域學期成績，作為科學性向測驗的效標，取得有效樣本比例 100%。

（四）區辨效度研究樣本

以高屏地區 30 所學校數理資優班的資優生，以及同校任一班級的普通生為樣本，進行測驗區辨效度之研究，取得資優生 533 人（男 370 人、女 163 人）；一般生 718 人（男 364 人、女 354 人），樣本有效比例為 99%。

（五）高中科學性向測驗常模樣本

為說明題庫的運用方式，並因應全國資優學生鑑定之實務現況，乃從高中科學性向題庫選題組卷，並建置常模，由於高中資優學生鑑定於 8 月舉行，考量國內高中學校仍有能力分校的現況，且並非所有學生皆進入高中職就讀，因此常模建置以臺灣地區九年級學生為母群體，依臺灣北中南東各區（未

表 4　國中科學性向測驗區域樣本

卷別	試別	第 1 次	第 2 次	第 3 次	第 4 次	第 5 次
	學校區域	南區	南區	南區	南區	中區
	學校數（學校規模大 / 中 / 小）	14(5/7/2)	13(3/8/2)	14(4/8/2)	10(3/5/2)	13(4/7/2)
A 卷		345				
B 卷		342				
C 卷		338				
D 卷		336				
E 卷			364			
F 卷			363			
G 卷			370			
H 卷				449		
I 卷				498		
J 卷				508		
K 卷					472	
L 卷					463	
M 卷					455	
N 卷						279
O 卷						285
P 卷						304
Q 卷						302
合計（有效比例）		1,361(94%)	1,097(91%)	1,455(95%)	1,390(99%)	1,170(98%)

表 5　高中科學性向測驗區域樣本

卷別	試別	第一次	第二次	第三次
	學校區域	南區	南區	中區
	學校數（學校規模大 / 中 / 小）	16(5/8/3)	13(3/8/2)	14(3/9/2)
A 卷		379		
B 卷		365		
C 卷		383		
D 卷			278	
E 卷			295	
F 卷			297	
G 卷				411
H 卷				392
I 卷				414
J 卷				355
合計（有效比例）		1,127(92%)	870(95%)	1,572(98%)

含離島）學生數比例，以分層叢集取樣方式選定 18 所學校，每校各一班，共 745 人，有效樣本比例為 99%，樣本取樣適切性考驗 KMO 值為 .92，屬於適切指標。

五、資料處理方式

本研究使用 SPSS 24.0 和 BILOG-MG 3.0 統計軟體進行以下分析：

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

2. 在 CTT 的基礎下，計算所有預試後試題之答對率、鑑別度、各題與總分相關以及選項誘答力。

3. 在 IRT 的基礎下，以等化的方式分析試題訊息，計算出試題鑑別度、難度和猜測度三參數。

4. 以獨立樣本 *t* 考驗，瞭解題庫試題在資優生與一般生的區辨效度。

5. 以積差相關，瞭解科學性向測驗與學校自然科學業成就之關聯效度。

研究結果

一、試題品質之確保

（一）依科學性向理論架構命題

研究先後執行 4 次命題工作，各次命題工作開始前，皆從人才資料庫中依學科專長遴選委員，總計有 72 位委員參與國中、高中科學性向試題命題。委員確認後先召開命題說明會議，確保命題委員充分瞭解本研究之科學性向理論架構內涵，並能依此進行命題，再者，也說明以及過程中應注意事項。除此之外，亦提供試題編寫手冊，供命題委員會議後參考，初步確保試題有好的內容效度。

（二）透過專家審查與學生試做來提升試題品質

本研究由物理、化學、生物、地科領域教授分科組成審查團隊，各科審查團隊由至少一名同時具備專業學科知識及科學教育知識背景的教授以及三至六名具備專業學科知識的教授組成。四個審查團隊成員穩定，六年來有極佳工作默契，審題時能充分就題幹敘述、圖表資訊、選項敘述與答案的正確性檢核，並做適度的修改，以降低命題者洩題的風險；也會就試題所屬的認知歷程、能力向度做確認。對於不適切的試題可逕予刪除或交由命題者修正後再次進行審查；對於合宜的試題，則由研究者與另一位同為測驗專長的教授，依選擇題命題原則，來檢視試題敘述的合宜性並做適當的修改。

除了仰賴專家的審查外，所有試題皆隨機組成學生能於 30 分鐘內完成的短測驗，由長期合作學校的六年級以及九年級學生試做，透過考生的觀點，提供試題的回饋，再者，所有施測過程皆由固定的主試人員實施，以利蒐集學生作答歷程的反應，讓試題品質更加完善。

二、校準題庫內試題訊息分析

本研究的試題將做為資優學生鑑定之用，考量鑑定辦法規範，須以全國學生的施測結果進行參照，因此以全國學生為樣本，先建立可作為未來試題校準依據的題庫，以利未來擴增試題時，可以僅作地區性的施測，再利用校準題庫內的試題進行定錨，以取得不受樣本影響的獨立試題參數。

研究者將通過品質分析後的試題，依學科授課時數比例進行組卷，採共同測驗隨機

組設計進行水平等化，兩份測驗之間有 8 題的連結試題（各科兩題），提供不同測驗的共同參照點，讓不同測驗即使施測不同的樣本群體，也能有基準相同的參數。

本研究以 IRT 分析試題的 a、b、c 三參數，參考每一試題的內適合度（infit）與外適合度（outfit）作為題目刪選的參考指標之一，適合度 *Z* 值大於 3 以及小於 -3，需加以檢討。除了參考適合度外，也採用傳統測驗理論進行題目分析，來瞭解四個選項的誘答力，如果選項的誘答力太低，或鑑別度出現負值，則重新交付審題委員進行修題，不納入校準試題題庫中。

（一）國中科學性向校準試題題庫

1. **試題構念分析**

表 6 是學科領域與認知能力向度雙向細目表，由表中可知題庫內，在學科領域方面物理有 55 題、化學有 32 題、生物有 35 題，及地球有 18 題；在認知能力向度方面，理解有 54 題、應用有 37 題、分析有 41 題及評鑑有 8 題。

表 7 是學科領域與科學能力向度雙向細目表，由表中可知，在科學能力向度方面，科學地解釋現象有 50 題、評估與設計科學探究有 20 題、科學地詮釋資料與證據有 38 題，及科學問題解決有 32 題。

2. **試題訊息分析**

經等化分析後，題庫內所有題目的鑑別度 a 參數平均值為 0.696，標準差 0.222；難度 b 參數平均值為 0.599，標準差 1.444；猜測度 c 參數平均值為 0.213，標準差 0.050。

測驗的題目皆為四選一選擇題，必然有某種程度的猜對率，目前 c 參數平均值 0.213 屬於適切地反應出此一合理的猜測度；再由 b 參數平均值 0.599 表示試題中間偏難，因為題庫建置目的是為了資優鑑定使用，未來擴增試題，宜持續增加難度參數高的題目；a 參數平均值 0.696，大於 0.6，表示試題平均

表 6　國中科學性向測驗校準試題題庫之學科領域與認知能力向度雙向細目表

學科領域	認知能力向度				題數
	理解	應用	分析	評鑑	
物理	13	18	21	3	55
化學	8	13	10	1	32
生物	24	2	8	1	35
地球科學	9	4	2	3	18
合計	54	37	41	8	140

表 7　國中科學性向測驗校準試題題庫之學科領域與科學能力向度雙向細目表

學科領域	科學能力向度				題數
	科學地解釋現象	評估與設計科學探究	科學地詮釋資料與證據	科學問題解決	
物理	9	12	12	22	55
化學	12	6	10	4	32
生物	20	2	11	2	35
地球科學	9	0	5	4	18
合計	50	20	38	32	140

鑑別度佳。

（二）高中科學性向校準試題題庫

1. **試題構念分析**

表 8 中可知校準題庫內，在學科領域方面包含物理 82 題、化學 67 題、生物 71 題及地球 54 題；在認知能力向度方面，理解能力有 141 題、應用能力有 44 題、分析能力有 66 題及評鑑能力有 23 題；再由表 9 可知在科學能力向度方面，科學地解釋現象有 121 題、評估與設計科學探究有 13 題、科學地詮釋資料與證據有 86 題及科學問題解決有 54 題。

2. **試題訊息分析**

經等化分析後，題庫內所有題目的鑑別度 a 參數平均值為 1.080，標準差 0.460；難度 b 參數平均值為 1.009，標準差為 1.100；猜測度 c 參數平均值為 0.226，標準差為 0.062。

c 參數平均值 0.226 屬於適切地反應出此一四選一選擇題合理的猜測度；由 b 參數平均值 1.009 看來，整體高中科學性向測驗題目，符合做為資優生甄選的目的；鑑別度 a 參數平均數為 1.080，表示題目鑑別度佳。

（三）單向度考驗結果

本研究採用的 IRT 模式假定所測量的能力具單一向度性，Lord（1980）認為，欲檢測測驗是否符合單一主要向度之要求時，如果題目之四分相關矩陣的主對角線是共同性估計值，即可用該相關矩陣的特徵值來評估。當第一特徵值明顯大於第二個，第二特徵值沒有明顯大於其他特徵值，那該測驗就近似單一向度。本研究的各式測驗皆符合此一假定，茲第一份全國施測的高中科學性向測驗的因素分析結果為例，其第一個特徵值為 5.480、第二個特徵值為 1.660、第三個特徵值為 1.628，第一特徵值明顯大於第二特徵值（約 3.3 倍），而第二特徵值則與第三特徵值相差無幾，可說明測驗為單一向度。

表 8　高中科學性向測驗校準試題題庫之學科領域與認知能力向度雙向細目表

學科領域	認知能力向度				題數
	理解	應用	分析	評鑑	
物理	50	12	6	14	82
化學	35	16	13	3	67
生物	33	8	27	3	71
地球科學	23	8	20	3	54
合計	141	44	66	23	274

表 9　高中科學性向測驗校準試題題庫之學科領域與科學能力向度雙向細目表

學科領域	科學能力向度				題數
	科學地解釋現象	評估與設計科學探究	科學地詮釋資料與證據	科學問題解決	
物理	23	3	24	32	82
化學	33	3	25	6	67
生物	40	5	23	3	71
地球科學	25	2	14	13	54
合計	121	13	86	54	274

三、定錨測驗試題分析

完成校準題庫後，未來擴充試題時，只要透過等化技術即可讓各次測驗分數都能在同一個量尺上，不受不同受測者的分數與不同次測驗影響，為了減少題庫共同試題的曝光程度，提高試題保密性，研究者從中篩選出一式涵蓋各科的定錨測驗，做為未來各式擴充測驗與校準題庫之間的共同測驗。

未來提供鑑定用的國中科學性向測驗設定為 34 題的測驗，因此以 20% 為定錨試題，故每卷篩選出 7 題組成定錨測驗，內容包含物理 2 題、化學 2 題、生物 2 題及地球科學 1 題；在認知能力向度中，分析能力有 2 題、應用能力有 2 題及理解能力有 3 題；在科學能力向度中，科學地解釋現象有 3 題、科學問題解決有 1 題及科學地詮釋資料與證據有 3 題。國中科學性向定錨測驗，鑑別度 a 參數平均值為 0.688，標準差為 0.240；難度 b 參數介於 -0.270 至 1.268 之間，平均值為 0.695，標準差為 0.369；猜測度 c 參數平均值為 0.184，標準差為 0.023。

在高中科學性向測驗，亦設定為 50 題的測驗，因此以 20% 為定錨試題，故每卷篩選出 10 題，包含物理有 3 題、化學有 2 題、生物有 3 題及地球科學有 2 題；在認知能力向度中，分析能力有 3 題、應用能力有 2 題、理解能力有 3 題及評鑑能力有 2 題；在科學能力向度中，科學問題解決有 4 題、科學地詮釋資料與證據有 4 題及評估與設計科學探究有 2 題。高中科學性向定錨測驗，鑑別度 a 參數平均值為 1.656，標準差為 0.441；難度 b 參數介於 0.112 至 1.359 之間，平均值為 0.848，標準差為 0.434；猜測度 c 參數平均值為 0.205，標準差為 0.063。

四、擴增的試題訊息分析

研究者將通過試題品質檢核的試題，分批組卷以區域樣本持續進行施測，並透過定錨測驗與校準試題題庫進行水平等化以取得試題的三參數。以下分別針對國中科學性向擴增試題與高中科學性向的擴增試題進行說明。

（一）國中科學性向試題

研究者在四年期間總計組成 17 卷國中科學性向測驗，分為五次進行施測。由表 10 及表 11 可知，通過統計分析可以放入題庫的試題共計有 309 題。其中在學科知識中，計有物理有 107 題、化學有 79 題、生物有 73 題及地球科學有 50 題；在認知能力向度中，理解能力有 110 題、應用能力有 72 題、分析能力有 108 題及評鑑能力有 19 題；在科學能力向度中，科學地解釋現象有 134 題、評估與設計科學探究有 33 題、科學地詮釋資料與證據有 88 題及科學問題解決有 54 題。

經與校準試題題庫等化後，擴增題目的鑑別度 a 參數平均值為 0.673，標準差為 0.237；難度 b 參數平均值為 1.265，標準差為 1.351；猜測度 c 參數平均值為 0.210，標準差為 0.052。

c 參數平均值屬於適切地反應出四選一選擇題合理的猜測度；至於 b 參數平均值 1.265，已相較於校準題庫，提升了試題難度，鑑別度 a 參數平均值亦適切，符合作為資優甄選測試的目的。

再進一步由表 12 可知，國中科學性向測驗擴增試題難易程度分布，b 參數在 -0.5 以下有 25 題，介於 -0.5 至 1.5 之間試題有 150 題，在 1.5 以上試題有 134 題。

（二）高中科學性向試題

研究者在三年期間總計組成 10 卷高中科學性向測驗，每卷 50 題試題，包含 10 題定錨試題，各卷作答時間設定為 70 分鐘，

表 10　國中科學性向測驗擴增試題之學科領域與認知能力向度雙向細目表

學科領域	認知能力向度				題數
	理解	應用	分析	評鑑	
物理	28	32	44	3	107
化學	15	21	33	10	79
生物	47	7	16	3	73
地球科學	20	12	15	3	50
合計	110	72	108	19	309

表 11　國中科學性向測驗擴增試題之學科領域與科學能力向度雙向細目表

學科領域	科學能力向度				題數
	科學地解釋現象	評估與設計科學探究	科學地詮釋資料與證據	科學問題解決	
物理	23	15	32	37	107
化學	37	10	22	10	79
生物	50	5	18	0	73
地球科學	24	3	16	7	50
合計	134	33	88	54	309

表 12　國中科學性向測驗擴增試題難易程度分布表

學科領域	b 參數			題數
	-0.5 以下	-0.5~1.5	1.5 以上	
物理	8	44	55	107
化學	8	44	27	79
生物	5	42	26	73
地科	4	20	26	50
合計	25	150	134	309

分為三次進行施測。由表 13 及表 14 可知，通過統計審查可以放入題庫的試題共計有 412 題。其中在學科知識中，物理有 117 題、化學有 111 題、生物有 102 題及地球科學有 82 題；在認知能力向度中，理解能力有 168 題、應用能力有 93 題、分析能力有 126 題及評鑑能力有 25 題；在科學能力向度中，科學地解釋現象有 188 題、評估與設計科學探究有 23 題、科學地詮釋資料與證據有 135 題及科學問題解決有 66 題。

經與校準試題題庫等化分析後，擴增題目的鑑別度 a 參數平均值為 1.190，標準差為 0.446；難度 b 參數平均值為 0.947，標準差為 0.936；猜測度 c 參數平均值為 0.235，標準差為 0.063。

c 參數平均值 0.235 適切地反應出合理的猜測度；鑑別度 a 參數亦佳；至於 b 參數平均值 0.947 符合作為資優甄選測試的目的。

再進一步由表 15 可知，高中科學性向測驗擴增試題難易程度分布，b 參數在 -0.5

表 13　高中科學性向測驗擴增試題之學科領域與認知能力向度雙向細目表

學科領域	認知能力向度				合計
	理解	應用	分析	評鑑	
物理	40	33	33	11	117
化學	48	27	30	6	111
生物	45	8	45	4	102
地球科學	35	25	18	4	82
合計	168	93	126	25	412

表 14　高中科學性向測驗擴增試題之學科領域與科學能力向度雙向細目表

學科領域	科學能力向度				合計
	科學地解釋現象	評估與設計科學探究	科學地詮釋資料與證據	科學問題解決	
物理	40	7	31	39	117
化學	49	12	41	9	111
生物	62	1	35	4	102
地球科學	37	3	28	14	82
合計	188	23	135	66	412

表 15　高中科學性向測驗擴增試題難易程度分布表

學科領域	難度（b 參數）			題數
	-0.5 以下	-0.5~1.5	1.5 以上	
物理	6	54	57	117
化學	5	55	51	111
生物	3	69	30	102
地球科學	7	34	41	82
合計	21	212	179	412

以下的試題有 21 題；b 參數介於 -0.5 至 1.5 之間試題有 212 題；b 參數在 1.5 以上試題有 179 題。

五、題庫試題區辨效度分析

為了瞭解題庫的試題是否具備良好的區辨效度，研究者隨機由兩個題庫挑選試題組卷進行施測，其中 14 題來自國中科學性向測驗題庫，18 題來自高中科學性向測驗題庫，並同時對八年級資優生與一般生施測。

由表 16 可知，資優生在測驗各向度的得分，都明顯高於一般生，獨立樣本 *t* 檢定分析結果，在所有向度皆達顯著水準（$p<.0025$），效果量介於 0.70 至 1.02 之間，呈現高度效果，可見科學性向試題有良好的區辨效度。

六、題庫試題關聯效度分析

本研究以三個班級的數理資優學生，在校自然科學業總成績，與科學性向測驗成

表 16　數理資優生與一般生在科學性向測驗之獨立樣本 *t* 檢定摘要表

分向度	題數	資優生（N=568）		一般生（N=718）		*t* 值	*p*	效果量（*d*）
		平均數	標準差	平均數	標準差			
認知歷程 - 分析	13	8.202	2.903	5.526	2.390	17.722	.000	1.018
認知歷程 - 理解	7	3.783	1.655	2.435	1.458	15.518	.000	.871
認知歷程 - 應用	9	5.558	2.171	3.708	1.921	16.195	.000	.909
認知歷程 - 評鑑	3	2.178	.894	1.543	.915	12.513	.000	.701
能力向度 - 設計	6	3.600	1.589	2.206	1.435	16.303	.000	.926
能力向度 - 詮釋	12	7.444	2.705	5.208	2.327	15.923	.000	.894
能力向度 - 現象	7	4.033	1.659	2.756	1.484	14.547	.000	.817
能力向度 - 解決	7	4.644	1.732	3.042	1.510	17.429	.000	.994

績進行相關係數分析，結果發現相關係數值介於 .47 至 .51 之間，*p* 值均達顯著水準（$p<.001$）；決定係數（R^2）介於 .22 至 .26 之間，符合預期之相關（如表 17）。

七、運用題庫組卷以供資優鑑定之用

題庫建立完成後，若因為鑑定需要測驗時，即可由題庫中選題組卷，以因應資優鑑定之用，以下乃列舉一例說明如何運用題庫。

（一）依試題統計訊息與能力向度組成測驗

研究者考量科目比例、各科目之認知能力，與科學能力向度比例，也參酌各試題的參數，共選取 50 題組成高中科學性向測驗。表 18 為此卷的學科領域與認知能力向度雙向細目表，包含物理有 15 題、化學有 10 題、生物有 15 題及地球科學有 10 題；在認知能力向度中，理解能力有 24 題、應用能力有 8 題、分析能力有 13 題及評鑑能力有 5 題；再由表 19 可知，在科學能力向度中，科學地解釋現象有 20 題、評估與設計科學探究有 2 題、科學地詮釋資料與證據有 14 題，及科學問題解決有 14 題。

（二）測驗訊息函數

表 17　自然科學業成績與性向測驗關聯效度分析

	A 班性向測驗	B 班性向測驗	C 班性向測驗
A 班學業成績 (N=29)	.473***(R^2=.22)		
B 班學業成績 (N=19)	-	.483***(R^2=.23)	
C 班學業成績 (N=28)	-	-	.511***(R^2=.26)

***$p < .001$，括號內為決定係數

表 18　高中科學性向測驗學科領域與認知能力向度雙向細目表

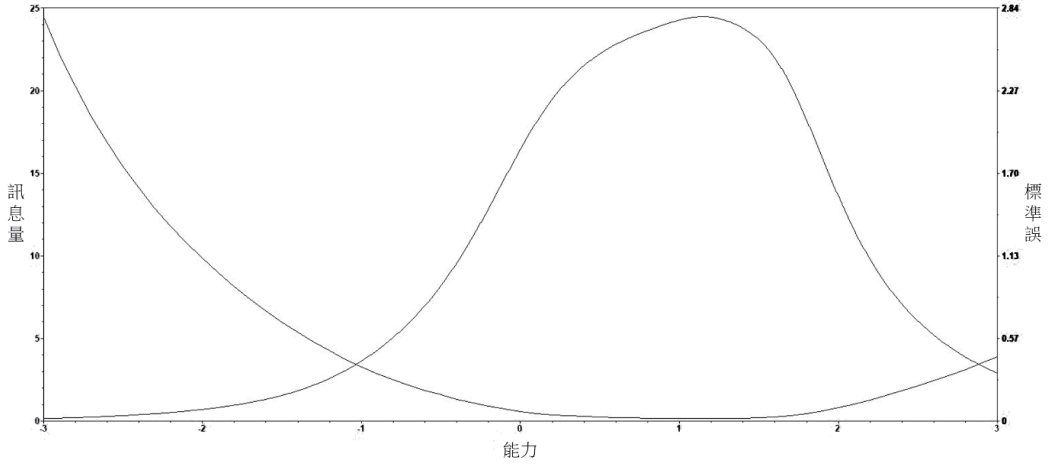
學科領域	認知能力向度				題數
	理解	應用	分析	評鑑	
物理	8	3	1	3	15
化學	3	5	2	0	10
生物	8	0	6	1	15
地球科學	5	0	4	1	10
小計	24	8	13	5	50

表 19　高中科學性向測驗學科領域與科學能力向度雙向細目表

學科領域	科學能力向度				題數
	科學地解釋現象	評估與設計科學探究	科學地詮釋資料與證據	科學問題解決	
物理	2	0	3	10	15
化學	4	1	4	1	10
生物	9	1	4	1	15
地球科學	5	0	3	2	10
小計	20	2	14	14	50

圖 5 是本測驗的測驗訊息函數（TIC）曲線圖，圖中呈現測驗的訊息量及能力估計值的標準誤。由圖可知，訊息量最高點及標準誤最低點，都在 +1.1 的地方，顯示測驗對能力在平均數以上 1.1 個標準差的學生，能提供最大的訊息及最小的誤差。

圖 5　自然科學性向測驗的 TIC 及能力估計值的標準誤



（三）常模

由於我國＜身心障礙及資賦優異學生鑑定辦法＞規定資優學生鑑定的標準為百分等級 97，或平均數以上 2 的標準差，因此，本測驗也提供同年齡學生的百分等級常模與常態轉換後的 T 分數常模（如表 20），供實務使用。

結論與建議

一、結論

（一）提出三向度的科學性向理論架構作為題目發展的依據

在文獻探討與多次專家小組討論後，科學性向採用的理論構念包含：學科知識、認知歷程與科學能力三向度。考量學術性向測驗有預測學生未來成就表現的功能，因此學科知識是以國中、高中自然領域課程的內涵做考量，區分為物理、化學、生物和地球科學四個科目來發展題目，在研究執行的後期，十二年國教的課程綱要公布並實施，為了呼應十二年國教的理念，在最後一輪命題時，也在四個學科之外增加跨學科的試題。

表 20　高中自然科性向測驗原始分數常模對照表

原始分數	百分等級常模	常態轉換 T 分數常模	原始分數	百分等級常模	常態轉換 T 分數常模
2	1 以下	17.93	25	55	51.35
3	1 以下	18.76	26	59	52.31
4	1 以下	19.59	27	63	53.22
5	1 以下	20.41	28	66	54.08
6	1 以下	21.24	29	69	54.88
7	1 以下	23.50	30	72	55.73
8	1	25.28	31	74	56.59
9	1	26.57	32	77	57.40
10	2	27.86	33	80	58.32
11	3	30.49	34	83	59.31
12	5	33.58	35	85	60.32
13	8	36.03	36	87	61.38
14	11	37.98	37	90	62.55
15	16	39.99	38	92	63.62
16	20	41.61	39	93	64.53
17	24	42.82	40	94	65.46
18	27	43.99	41	95	66.61
19	32	45.21	42	97	67.96
20	37	46.55	43	98	69.63
21	41	47.69	44	99	71.76
22	44	48.55	45	99	73.50
23	48	49.41	46	99 以上	74.38
24	51	50.34	47	99 以上	76.50

認知歷程以高層次思考能力為主，包含理解、應用、分析和評鑑四個認知能力。科學能力則考量科學學習的能力，提出科學地解釋現象、評估與設計科學探究、科學地詮釋資料與證據與科學問題解決四種科學能力。

（二）題庫試題皆依據架構命題並嚴謹的審查，確保品質

題庫中試題由許多命題委員於不同時間命題，為了確保試題的內容效度與品質，所有題目皆先交付物理、化學、生物、地科專家分科組成的修審小組，檢核題目與命題架構符合狀況，確認題幹敘述、圖表資訊、選項敘述與答案的正確性，並對題目做適度的修改，一方面以降低原始命題者洩題的風險；二方面對於可能涉及的學科基礎知識，則摒除所有專有名詞，以淺顯的字詞來詮釋其意義，以免能力優秀學生受困於名詞理解，反而無法於測驗之中反應其優秀潛能。學科知識內容確認後，則由測驗專家就試題編製原則再予審查，也透過小樣本的實地測試，取得學生實際反應資料作為題目修審的參考。本研究以嚴謹的程序，確保題庫中的題目皆有良好的品質。

（三）以 IRT 的測驗發展步驟建立適用於我國目前的資優學生鑑定之題庫

由於一個試題充足的優良題庫並非一蹴可及，因應教育部（2007）要求資優生鑑定應以全國常模作參照，因此本研究先分別建立國中科學性向測驗以及高中科學性向測驗的全國代表性樣本的學生能力參數以及試題參數未來可從中選擇題目作為已校準題目，採用題目預先等化設計方式，透過共同測驗，使不同試卷的題目參數，化為相同量尺單位，擴充題庫規模。由於國中科學性向測驗的全國樣本規模有 4,663 人，高中科學性向測驗的全國樣本規模有 3,702 人，未來可以此作為校準題庫，讓未來新題目的施測可以僅選擇區域型樣本即可，以節省人力等相關成本，也有利於取得學校的支持，裨益於題庫的擴充。

（四）每學期第一次成績考查結束後是最佳的施測時間

計畫執行過程為確保試題保密性，必須由受過訓練的主試人員親自施測，再者，學校端也須配合聯繫取樣班級師生，由於本研究取樣對象皆為即將畢業年級，為了提高學校的合作意願，研究者發現每學期第一次成績考查後是最佳的施測時間，研究團隊可於寒、暑假與學校聯繫，請學校安排新學期開學後 3-4 月及 9-10 月的時間實施測驗，這段期間亦可作為知情同意書發送的作業時間，此一題庫擴充工作能永續落實。

（五）由校準題庫中提取一組固定的定錨試題以降低題目外洩的風險

題庫建置需要不斷地擴充試題數量，為了讓各次測驗分數都能在同一個量尺上，不受不同受測者的分數與不同次測驗影響，透過共同測驗來進行測驗之間的等化，本研究為了減少共同試題的曝光程度，提高題庫試題保密性，乃分別由國中科學性向測驗題庫與高中科學性向測驗各篩選出一式涵蓋各科的定錨測驗，做為未來各式擴充測驗與校準題庫之間的共同測驗，二式定錨測驗不僅考量性向測驗各構念配比，三參數的數值也理想。

（六）具有良好的關聯效度與區辨效度

由題庫中選題組成的測驗，與取樣學校資優學生的自然領域學期成績有中度相關，符合性向測驗與目前成就的關係期待；而在資優生與一般生的施測結果，資優生在各題的答對率都顯著高於一般生，效果量 *d* 介於 0.701 至 1.018 之間，可見科學性向題庫試題

有良好的區辨效度。

（七）實測組卷的品質良好，支持題庫的效用性

本研究由題庫中組成一卷測驗建置百分等級及常態化的 T 分數常模，供實際鑑定考試之用。測驗以對能力在 +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，經與 745 人的常模資料交互驗證，結果測驗對於高能力者能提供精確的能力評估。

二、建議

（一）持續擴充資優學生學術性向測驗題庫

本研究已建立了有全國性樣本的校準試題，提供所有試題 IRT 的題目參數，建議未來可以採用題目參數連結的方式，利用定錨測驗持續來擴充資優學生學術性向測驗題庫，當題庫中試題的數量夠多，就能降低對測驗題目外流的擔憂（Vale, 2006），也能有助於組成更有品質的測驗。

（二）可考慮開發電腦化適性測驗

因應數理學術性向資優學生的鑑定，目前國中學生科學性向測驗與高中學生科學性向測驗，每年皆會辦理，且參與測驗的學生也相當多。近兩年受到 COVID 19 疫情的干擾因素，讓鑑定辦理單位對於在群聚空間的測驗能否辦理？對於未能如期參加測驗的居家隔離考生如何補測？平行測驗不足？都有許多的困擾，本研究採用 IRT 來進行題庫的建置，所有題庫中試題皆有鑑別度、難度和猜測度三參數，未來可以考慮進一步開發電腦化適性測驗，來克服上述的問題。

（三）可在增加跨領域的試題

目前題庫試題在學科領域部分，分成物理，化學、生物與地球科學四科目，考量目前十二年國教課綱重視跨領域的學習，未來可增加跨領域的試題。

（四）未來可以依據訊息量，選題組成合適的試卷

題庫的題目發展與單份測驗題目發展最大的不同在於，題庫發展主要關注發展高品質的題目即可，題目之間是否有重複概念、題目與全測驗的關聯性都可於組測驗卷時再做考量。在本文中，受限於篇幅，故僅選擇高中科學性向測驗題庫來說明如何組測驗卷運用，由於全國資優學生鑑定，需遵從 < 身心障礙及資賦優異學生鑑定辦法 > 之規定，以全部學生為母群建立常模，採 PR97 或平均數以上 2 標準差進行切截，因此研究者乃示範以 +1 以上作為測驗訊息函數，一方面考量科目比例、各科目之認知能力，與科學能力向度比例，另一方面也參酌各試題的參數，進行組卷，再建置和於目前法規規範鑑定所需的常模。另一國中科學性向題庫，亦可以相同的方式來組卷應用。

易言之，由於題庫內的所有試題都有其鑑別度、難度和猜測度三參數，未來使用可以考量施測目的與需要，自訂測驗訊息函數，參酌各試題的參數，選擇適切的題目組成測驗卷，例如，用於資優鑑定的試卷可以設定介於 +1 至 +2 的測驗訊息函數，來由題庫選取適合的試題。不會受題庫整體難度影響資優學生的鑑出率。

參考文獻

余民寧（2009）：試題反應理論（IRT）及其應用（初版）。心理。[Yu, M.-N. (2009). *Item response theory* (1st). Psychological.]

余民寧（2011）：教育測驗與評量：成就測

驗與教學評量（三版）。心理。[Yu, M.-N. (2011). *Educational testing and assessment* (3rd). Psychological.]

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（未出版碩士論文）。國立臺灣師範大學。[Lee, B.-S. (2003). *Research on major choice and prediction about academic achievements for the first grade of senior high students* (Unpublished Master's thesis). National Taiwan Normal University.]

身心障礙及資賦優異學生鑑定辦法（2013）：中華民國一百零二年九月二日教育部臺教學（四）字第 1020125519B 號令修正發布。[Ministry of Education (2013). Identification Methods for Students with Disabilities and Gifted Talents.]

侯雅齡（2010）：「國民中學自然科學性向測驗」之編製。測驗學刊，57（1），29-58。[Hou, Y.-L. (2010). The development of natural science academic aptitude tests for junior high school students. *Psychological Testing, 57*(1), 29-58.] https://doi.org/10.7108/PT.201003.0029

侯雅齡（2013）：資優生科學自我概念與科學成就之縱貫研究。教育科學研究期刊，58（2），57-90。[Hou, Y.-L. (2013). Longitudinal study of gifted students' science self-concept and science achievement. *Journal of Research in Education Sciences, 58*(2), 57-90.] https://doi.org/10.3966/2073753X2013065802003

侯雅齡、洪振方、吳裕益、古源光（2018）：研訂編製題庫化資優鑑定工具規範暨建立國高中自然科性向測驗題庫。教育部國民及學前教育署委託之研究成果報告（編號：C10430），未出版。[Hou, Y.-L., Hung, J.-F., Wu, Y.-Y., & Gu, Y.-G. (2018). *Research and formulate the specification of gifted identification tools based on item bank and establish the item bank of natural science subject aptitude test in junior high school.* K-12 Education Administration, MOE, commissioned research report (C10430).]

特殊教育法（2019）：中華民國一百零八年四月二十四日總統華總一義字第 10800039361 號令修正公布。[Ministry of Education (2019). The Special Education Act.]

高級中等學校適性學習社區教育資源均質化實施方案（2017）：中華民國一百零六年十一月十日臺教授國部字第 1060090835 號函修定。[K-12 Education Administration. (2017/11/10). Implementation plan for homogenization of educational resources in community in senior high schools adaptive education (No. 1060090835).]

張正芬（2014）：身心障礙及資賦優異學生鑑定辦法說明手冊（國立臺灣師範大學特殊教育學系）。教育部。[Zhang, Z.-F. (2014). *Handbook of Identification Methods for Students with Disabilities and Gifted Talents* (Department of Special Education, National Taiwan Normal University). Ministry of Education.]

教育部（2007 年 10 月 17 日）：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。[Ministry of Education. (2007, October 17). *Revise the legality checklist of the county and city government's gifted student identification process.*]

郭伯臣、許天維、黃志傑、曾玉琳（2003）：定錨試題分佈對測驗等化效果之影響。測驗統計年刊，11，67-92。[Kou, B.-C., Sheu, T.-W., Huang, C.-C., & Tseng, Y.-L. (2003). Equating performances on different anchor item distributions. *Journal of Research on Measurement and Statistics, 11*, 67-92.] https://doi.org/10.6773/JRMS.200312.0067

綜合報導（2017年8月22日）：高中資優班考試傳洩題 補習班一字不差16題猜中12題。自由時報。[Liberty Times. (2017, August 22). *High school gifted and gifted class exam leaked questions, the cram school guessed 12 of the 16 questions verbatim*. Liberty Times.]

Anderson, L. W., & Krathwohl, D. R. (2001). *A taxonomy for learning, teaching, and assessing: A revison of Bloom's educational objectives*. Longman.

Bloom, B. S., Engelhart, M. D., Furst, E. J., Hill, W. H., & Krathwohl, D. R. (1956). *Taxonomy of educational objective*. In Taxonomy of educational objectives. David McKay.

Bramley, T., Crisp, V., & Shaw, S. (2019). The art of test construction: Can you make a good Physics exam by selecting questions from a bank? *Research Matters: A Cambridge Assessment Publication, 27*, 2-8. http://www.cambridgeassessment.org.uk/research-matters

Byrne, C. J. (1976). Computerized question banking systems: I - The state of the art. *British Journal of Educational Technology, 7*(2), 44-64. https://doi.org/10.1111/j.1467-8535.1976.tb00355.x

Crisp, V., Shaw, S., & Bramley, T. (2020). Should we be banking on it? Exploring potential issues in the use of 'item' banking with structured examination questions. *Assessment in Education: Principles, Policy and Practice, 27*(6), 655-669. https://doi.org/10.1080/0969594X.2020.1827220

Dorans, N. J., & Holland, P. W. (2000). Population invariance and the equatability of tests: Basic theory and the linear case. *Journal of Educational Measurement, 37*(4), 281-306. https://doi.org/10.1111/j.1745-3984.2000.tb01088.x

Dorans, N. J., Pommerich, M., & Holland, P. W. (2007). *Linking and aligning scores and scales*. Springer. https://doi.org/10.1007/978-0-387-49771-6

Hambleton, R. K., & Swaminathan, H. (1985). Assumptions of Item Response Theory. In R. K. Hambleton, & H. Swaminathan (Eds.), *Item Response Theory: Principles and Applications* (pp. 15-31). Springer. https://doi.org/10.1007/978-94-017-1988-9_2

Kolen, M. J., Brennan, R. L., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices*. Springer.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems* (1st ed.). Routledge.

Miller, D. M., Linn, R. L., & Gronlund, N. E., (2013). *Measurement and assessment in teaching*. (11th ed.). Pearson.

Malak, M. A. E. (2013). Question banking for English communication skills courses: A

guide. *European Scientific Journal, 9*(29), 458-477. http://citeseerx.ist.psu.edu/viewdoc/download?%0Adoi=10.1.1.824.1591&rep=rep1&type=pdf

Vale, D. C. (2006). Computerized item banking. In S. M. Downing, & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 276-300). Lawrence Erlbaum Associates, Inc., Publishers. https://doi.org/10.4324/9780203874776.ch11

Wood, R. (1968). The place and value of item banking. *Educational Research, 10*(2), 114-125. https://doi.org/10.1080/0013188680100204

Wright, B. D., & Bell, S. R. (1984). Item banks: What, why, how. *Journal of Educational Measurement, 21*(4), 331-345. http://www.jstor.org/stable/1434585 https://doi.org/10.1111/j.1745-3984.1984.tb01038.x

Zaaiman, H., van der Flier, H., & 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. https://doi.org/10.1023/A:1004020305088

收稿日期：2022.04.25
接受日期：2023.01.09

Bulletin of Special Education
2023, 48(1), 97-128
DOI: 10.6172/BSE.202303_48(1).0004

Development of item banks for middle school and high school science aptitude tests

Ya-Ling Hou
Professor,
Dept. of Special Education,
National Pingtung University of Education

Abstract

Rationale & Purpose: Because of the need to evaluate gifted students in scientific aptitude yearly in Taiwan, several methods are necessary to reduce the risk of leaked questions. The Programme for International Student Assessment, commissioned by Organization for Economic Co-operation and Development, emphasizes literacy in test design. One of the main goals of aptitude tests is to assess students' ability to navigate a rapidly changing society. Measuring scientific aptitude includes evaluating the ability to explain phenomena scientifically, evaluate and design scientific enquiries, solve scientific problems, and interpret scientific data and evidence. Scientific aptitude has considerable explanatory power for understanding the academic potential and learning attitudes of students. This study involved the construction of sustainable development item banks or question bank for the Science Aptitude Test taken by junior and senior high school students. Additional content can be added to the item banks, and the items we developed were constructed based on the same scale according to item response theory (IRT) for comparison. Consequently, when gifted students are identified, the appropriate items can be selected from the item bank for test design, and the item bank can be increased through proper way in the future. Developing a gifted identification item bank is a good practice, as it addresses the needs for such tools, appropriateness in the giftedness identification process, the necessity of maintaining fairness and rigor throughout the identification process, as well as the cost savings from avoiding duplicate developing test items. **Methods:** To determine the items for the Science Aptitude Test, we adopted three constructs: content knowledge, cognitive process, and scientific competency. All the items underwent strict content review, and the item banks were

developed using horizontal equalization based on IRT. Furthermore, every item had three parameters: parameter A was discrimination, parameter B was difficulty, and parameter C was guessing. In order to expand the Science Aptitude Test item bank for junior high school students, 4,663 students across Taiwan were analyzed to establish a 140-item calibration test. Thereafter, 7 of the 140 items were selected for the anchored test. One anchored test (seven items) and several new tests (each composed of 27 different items) were combined to form new tests. In total, 309 items were added. The three parameters for each item demonstrated good fit. For the high school Science Aptitude Test item bank, 3,702 high school students across the country participated to establish a 274-item calibration test. Only 10 of the 274 items were selected for the anchored test. One anchored test (10 items) and several new tests (each composed of 40 different items) were then combined to form new tests. In total, 412 items were added. Similarly, the three parameters demonstrated good fit. **Results/Findings:** The assignment committee members were selected using the talent database on the basis of their subject of expertise to ensure that the test items had high validity. A total of 72 committee members participated in designing the assignments for the scientific aptitude tests. The correctness and appropriateness of items were reviewed by four separate review teams with backgrounds in physics, chemistry, biology, and geology. Statistical tests revealed that the parameters for the anchor tests and aptitude tests (discrimination, difficulty, and guessing) all had good fit, and proved the items are unidimensional. Thereafter, high school students were tested, and the correlation between their aptitude test scores in the item bank and their semester scores in natural sciences indicated high criterion related validity. The test information curve indicated that the tests provide the maximum amount of information with the minimum number of errors when the student's capability is 1.1 standard deviations above the mean. Moreover, significant differences were observed between gifted students and other students in each item of the aptitude test, which further indicated good discriminant validity. Finally, 754 ninth-grade students were used as the norm to establish a percentile grade and a normalized t score norm through a 50-item test that selected from the item bank. **Conclusions/Implications:** After the literature review and several expert panel discussions, we adopted the three aforementioned constructs for scientific aptitude (content knowledge, cognitive process, and scientific competency). To reflect the 12-year national education system, interdisciplinary test questions were also added in the final round in addition to the four subjects. The test questions in the question bank were determined by many test committee members at different times. To ensure the content validity and quality of the test questions, all the questions were first given to the review team to determine whether the questions were consistent with the

test structure and make appropriate revisions accordingly. After the content knowledge revisions, the test experts reviewed the principles for compiling the test question. The response data of students from the fieldwork samples were referenced during revisions. This study used rigorous procedures to ensure that the questions in the question bank were of high quality. Because the Ministry of Education requires that the identification of gifted students must be based on national norms, this study first established national representative samples of the middle school and high school Science Aptitude Test. To allow future test committees to save money by conducting regional tests, labor and other related costs were conducive to obtaining the support of schools and expanding the question bank. Those managing the question bank must continually increase the number of test questions. To allow the scores of each test to be on the same scale and remain unaffected by the scores of other testers and tests, the equalization between tests was ensured using a common test. To reduce the exposure of the common test questions and improve the confidentiality of the test questions in the test bank, the researchers of this study selected a set of anchor tests covering each subject from the middle school and high school Science Aptitude Test to make additional tests and calibrations in the future. With the addition of the common test, the two-form anchor test not only considers the validity of the aptitude test constructs but also ensures that the values of the three parameters (Discrimination, Difficulty, and Guessing) are ideal.

Keywords: gifted students, item bank, item response theory, natural science academic aptitude test, test development