利用 Rasch 模式評估高中物理素養試卷過程：教師發展素養試題之啟示

林淑楞*、張仁壽、吳隆枝、黃建彰、趙臨軒、許程迪、
曾博淵、賴彥良、李俊穎

十二年國民基本教育將核心素養視為本次課程改革的重點，評量與課程目標一致才能有效推動改革。本文目的為介紹教育部國民及學前教育署普通型高級中等學校物理學科中心試題組教師與科教研究者共同合作發展與效化物理素養試卷後，運用 Rasch 模式中多向度部分給分模式，檢驗試題的品質、判讀 Wright map 的意義，呈現發展和施測物理素養試題所遇到的問題與解決策略，作為未來發展素養試題之參考。試題組團隊挑選 42 題物理素養試題，形成難易度相當的 A、B 兩卷。利用分層與便利取樣，向全臺 19 校，1119 位高三學生進行施測。Rasch 分析結果呈現兩卷幾乎所有試題與多向度部分模式都能適配，且兩卷各向度的平均難度值均在 0 附近。但 Wright map（能力與難度對應圖）呈現少數題目偏易，且缺乏高難度試題，恐較難有效評量高能力的學生。此研究在實務上建議建立全國統一的科學素養評量架構，未來並應多發展建構反應試題、以評量「評估及設計探究能力」和高難度的試題，並且善用 Wright map 增加教師對素養試題難度的敏感度並增進對學生能力的理解。評量設計探究能力和高難度的試題，並且善用 Wright map 增加教師對素養試題難度的敏感度與學生能力的了解。

關鍵詞：Rasch 模式、物理素養、評量、試題反應理論

* 林淑楞：國立彰化師範大學科學教育研究所教授兼所長
（通訊作者：sflin@cc.ncue.edu.tw）
張仁壽：國立基隆女中教師
吳隆枝：國立臺南二中教師
黃建彰：國立岡山高中教師
趙臨軒：國立鳳新高中教師
許程迪：高雄市立高雄中學退休教師
曾博淵：新北市立三重高中教師
賴彥良：國立嘉義高工教師
李俊穎：臺中市立中港高中教師

Journal of Educational Practice and Research
Vol. 34, No. 1, 121-154, Jun. 2021

Using Rasch Model to Evaluate Physics Competence Tests and its Implications for Teachers' Development of Competency-based Items

Shu-Fen Lin*, Jen-Shou Chang, Lung-Chih Wu, Chien-Chang Huang, Ling-Hsien Chao, Chen-Din Sheu, Po-Yuan Tseng, Yan-Liang Lai, & Chun-Ying Lee

Core competencies are the focus of the 12-year basic education curriculum reform. The purposes of the study were to introduce the development and validation of two physics competence tests through the cooperation of the seed teachers in the physics curriculum center, K-12 Education Administration, Ministry of Education, with a science education researcher, and to investigate strategies for developing competency items based on the analysis of multidimensional partial credit model, one of the Rasch models, and the interpretation of Wright maps. Forty-two physics competence items were selected and divided into test A and test B with a similar difficulty level. Stratified and convenience sampling was adopted. Two tests were administered to 1119 Grade 12 students from 19 public senior high schools in Taiwan. The results showed that most of the items in the two tests fit the multidimensional partial credit model, and the average of item difficulty for test A and test B were approximately 0. However, based on the Wright maps, some items were easier, and there was a lack of items with higher difficulty to assess students with higher ability. The study intends to inspire practical educators to build a national assessment framework for scientific literacy and to develop constructed-response items, items assessing the competence of evaluating and designing scientific inquiry, and items with high difficulty. The use of Wright maps may enhance teachers' sensitivity to item difficulty and their understanding of students' ability.

Keywords: ***assessment, Item Response Theory, physics competence, Rasch model***

* Shu-Fen Lin: Professor and Director, Graduate Institute of Science Education, National Changhua University of Education (corresponding author: sflin@cc.ncue.edu.tw)
Jen-Shou Chang: Teacher, National Keelung Girls' Senior High School
Lung-Chih Wu: Teacher, National Tainan Second Senior High School
Chien-Chang Huang: Teacher, National Kangshan Senior High School
Ling-Hsien Chao: Teacher, Feng-Hsin Senior High School
Chen-Din Sheu: Retired Teacher, Kaohsiung Municipal Kaohsiung Senior High School
Po-Yuan Tseng: Teacher, New Taipei Municipal SanChong High School
Yan-Liang Lai: Teacher, National Chia-Yi Industrial Vocational High School
Chun-Ying Lee: Teacher, Taichung Municipal Chung Gang Senior High School

利用 Rasch 模式評估高中物理素養試卷過程：教師發展素養試題之啟示

林淑梹、張仁壽、吳隆枝、黃建彰、趙臨軒、許程迪、
曾博淵、賴彥良、李俊穎

壹、前言

自從我國 2006 年參與國際學生能力評量計畫（Programme for International Student Assessment, PISA）後，以情境為題幹的類 PISA 試題便開始受到科學教育者的關注。由於 PISA 所評量的科學素養，主要是評量學生在模擬生活情境下，應用科學知識，展現解決問題的科學能力與態度（Organisation for Economic Co-operation and Development [OECD], 2017）。此目標與「十二年國民基本教育自然科學領域課程綱要」（以下簡稱 108 自然領綱）中兼重以科學認知為主的「學習內容」，與包含探究能力和科學的態度和本質的「學習表現」方向一致。另外，提升學生的核心素養是 108 課綱的教育目標，為了評量自然科學領域的核心素養，大學入學考試中心（以下簡稱大考中心）指出 111 學年度大學入學考試的命題，將配合 108 課綱強調核心素養和跨領域的精神，發展素養導向的試題（大學入學考試中心，2019）。教育部國民及學前教育署普通型高級中等學校物理學科中心（以下簡稱物理學科中心）自 105 學年起便開始培育高中種子教師發展評量物理素養的試題。本文旨在介紹高中物理教師和科教研究者初次合作發展和效化物理素養試卷的結果、Rasch 多向度部分給分模式用於檢測物理素養試題的功能，以做為未來教育工作者推動發展與效化物理或科學素養試卷之參考。

貳、背景說明

一、物理素養評量架構

近年來許多國家受到 PISA 評量科學、數學和閱讀素養的方式所影響，開始以人們可能接觸到的情境發展試題，評量受試者能否運用各領域的知識，解決問題，展現他們的科學、數學和閱讀能力。試題通常以情境為題幹，以單選題、簡答題、建構反應試題等多元題型構成題組，較過去以選擇題為主的測驗更能展現受試者真正的能力（McCoubrie, 2004），頗受到推動教育改革者所推崇。因此，基北區特色招生原本於 2013 年欲將「類 PISA 題型」納入數學和閱讀理解的素養評量，卻受到不少家長、教師和學生的反對聲浪下喊停（盧姮倩，2013 年 4 月 18 日）。為了將核心素養成為國民教育的主軸，我國於 2018 年公布 108 自然領綱，將培養未來公民科學素養訂定為當前課程改革的目標。大考中心將於 111 學年度大學入學考試和國中會考中，融入素養導向的試題（大學入學考試中心，2019）。總之，素養導向試題的發展已成為未來教育評量的趨勢。物理學科中心依循此趨勢，於 2016 年開始由種子教師研發物理素養試題。

108 自然領綱將課程架構分為「學習內容」和「學習表現」兩個部分，其中學習內容即各年段在自然科學領域四個科目的內容知識，學習表現則分為科學認知、探究能力和科學態度與本質三大部分，也就是認知、技能、情意三大面向。其中科學認知是指對內容知識的六大認知層次表現，也就是記憶、了解、應用、分析、評鑑和創造六個層次。探究能力包括思考智能和問題解決兩大部分：思考智能意旨探究過程中動腦想（minds-on）的各種能力，包括想像創造、推理論證、批判思辨和建立模型；問題解決則偏重探究過程中動手做（hands-on）的各種能力，包括觀察與定題、計劃與執行、分析與發現、討論與傳達（教育部，2018）。林蓓伶、潘昌志、蘇少祖與陳柏熹（2018）依據 108 自然領綱，建立科學素養導向題型命題參考架構，將「學習表現」為素養題型的評量目標，分別將核心概念分為三個層次：知曉、理解和應用；探究能力分為六項能力：(1)審視資訊並界定問題；(2)規劃實驗／探究計畫；(3)執行實驗／探究計畫；(4)分析與陳述數據；(5)解釋數據並說明推理過程；(6)分享與評鑑實驗／

探究結論；科學的態度與本質分為五個部分：(1)覺察；(2)動機與興趣；(3)審視與評價；(4)參與科學議題；(5)建立科學自我效能感。再利用「學習表現」三大面向繪製出三個雙向細目表，發展科學素養試題。另外選擇日常生活或學術探究模擬各種真實情境，以發展試題。

PISA 評量十五歲學生的科學素養，乃是此階段學生可能接觸的生活情境，發展試題，評量學生應用科學知識，以展現其解決問題的能力。PISA2015 科學素養評量架構中的情境包含個人、地區和全球的情境，應用的科學知識包含內容知識、程序性知識或認識論知識。科學能力則包含「解釋科學現象」（以下簡稱「解釋」）、「評估及設計科學探究」（以下簡稱「設計」）、「詮釋科學數據與證據」（以下簡稱「詮釋」）三種能力（OECD, 2017）。解釋面向能夠對自然與科技現象確認、提出和評估解釋，便需具備回憶和應用科學知識、提出解釋的模型、做出預測或假設等能力；設計面向則能夠區辨、確認科學問題，並能提出方法回答科學問題，評估不同方法的優劣等；詮釋面向則能分析和評估科學數據，利用各種表徵呈現主張與論點，並做出結論等。其中，設計和詮釋面向涵括的細項能力即是科學探究所需的能力，與 108 課綱中探究能力的學習表現內涵，以及林蓓伶等人（2018）所提出的六項探究能力一致。每一個試題乃評量受試者應用一種科學知識，以展現一種科學能力。由於 PISA2015 科學素養評量係採用電腦施測，而 PISA2006 採用紙筆測驗，試題型式稍有不同。物理學科中心考量現階段全國性考試仍採用紙筆測驗，因此參考 PISA2006 的題型，包括多重是非題、單選題、簡答題或建構反應試題（即開放性試題）四類（Bybee, McCrae, & Laurie, 2009）。Kind（2013）認為 PISA 以情境發展問題，利用這些試題確認學生是否有能力處理日常生活的問題，已跳脫評量學生解釋科學原理的方式。這些情境問題可讓學生展現出解決問題、進行科學研究、運用證據支持主張等能力。

美國科學教育標準對科學素養的定義為「個人對日常經驗的事物會因好奇而提問、尋求解答。個人具有描述、解釋、預測自然現象的能力。而且能夠閱讀、理解報章雜誌上有關科學的文章，能夠使用科學語言進行溝通」（National Research Council [NRC], 1996, p.22）。顯然，科學教育標準中所指的科學素養著重於科學能力的展現。Bybee（1997）提出科學素養的四個層級，以呈現人們科學素養的高低程度。這四個層級包括名義的（nominal）、功能的、概念與程序的（conceptual and procedural），以及多重面向的（multidimensional）科學素養。名義的科學素養意指學生只能認得一些與科學有關的概念；功能的科學素養意指學生能使用有限的理解來描述科學概念；概

念與程序的科學素養則指學生能理解科學中的知識結構、科學探究的過程和科技的設計；多重面向的科學素養展現學生不僅要了解概念性知識和程序性知識，也需要理解科學與科技面向中歷史、哲學和社會中與科學有關的事，即對科學本質的理解。運用各國課程的領域素養架構發展物理學科的評量試題，為物理教育學者普遍的依據（Adeleke & Joshua, 2015）。近年來，在物理教育中有些學者強調具有物理素養的學生能夠運用物理的知識透過說和寫解釋日常生活中的各種現象，以及運用數學和概念化的思考，分析和解決生活中有關科學的問題（Hurley & Henry, 2015; Körhasan & Gürel, 2019）。陳泰然（2015，頁 32）認為 108 自然領綱中高中必修課程應「加強微觀、抽象基本運算與理論推導理解科學；在教師的協助下，經歷完整科學探究之學習歷程；綜合運用所學，進行觀測、資料蒐集、分析歸納、解釋結論，以及撰寫完整科學報告。」由此可知，108 自然領綱期望學生至少須達到概念與程序的科學素養，其與 PISA2015 科學素養的評量架構（OECD, 2017）的思維相似，即利用內容知識、程序性知識和認識論知識展現科學能力（如圖 1）。因此，本研究採用 PISA2015 科學素養評量架構，作為發展物理素養試題的依據。其中，內容知識為 108 自然領綱中高中物理必修課程需學習的核心概念，而程序性知識則參考 PISA2015 科學素養評量架構所列舉之知識（如表 1）。

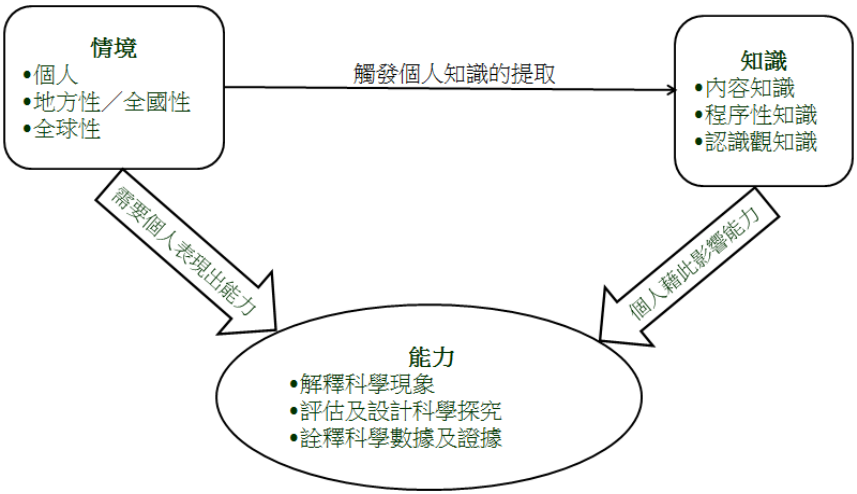


圖 1　以情境問題評量學生運用物理知識，展現科學能力。

表 1　物理素養評量架構中的內容知識與程序性知識

內容知識	程序性知識
● 能量的形式與轉換 ● 溫度與熱量 ● 自然界的尺度與單位 ● 力與運動 ● 波動、光及聲音 ● 萬有引力 ● 電磁現象 ● 量子現象 ● 基本交互作用 ● 科學在生活中的應用 ● 能源的開發與利用	● 變因的概念 ● 測量的概念 ● 測驗與減少不確定性的方法 ● 確保可複製與數據精確性之機制 ● 變因的控制策略及其在實驗設計中的角色，或利用隨機對照試驗，避免結果混淆或找出可能的因果機制。 ● 對已知科學問題知道設計合適的探究方法 ● 將數據抽象化或呈現的常用方法，如表格、圖形……。

二、試題反應理論在測驗上的應用

為了檢驗測驗工具的品質，試題發展者一般會採用古典測驗理論（Classical Test Theory, CTT）或試題反應理論（Item Response Theory, IRT）系統化地解釋測驗資料間的關係。其中，CTT 奠基於簡單線性函數假設（X= T＋E，X：可觀察分數，T：真實分數，E：誤差分數），或稱為真實分數模式。因計算公式簡單易懂，適用於大多數的教育與心理測驗和社會科學的資料分析。但是，CTT 有許多先天的限制，以致測驗理論學者們發展出 IRT 克服 CTT 的缺失（余民寧，2009）。雖然，IRT 所採用的計算公式較為複雜，模式不易理解，且須仰賴電腦軟體分析，對大多數試題發展者可能較為困難。然而，IRT 比 CTT 具備更多優點，近年來已有越來越多大型試題分析，如 PISA、美國學術評量測驗（Scholastic Assessment Test），或試題發展（陳彥君、洪素蘋，2019；張貴琳，2013；Kuo, Wu, Jen, & Hsu, 2015）採用 IRT 進行試題分析。以下簡介 IRT 克服 CTT 限制的特點，以增進試題發展者對 IRT 的認識。

IRT 是以非線性數學模式為基礎，受試者在一個試題的表現和受試者具備的某種能力具有一種非線性關係，這種非線性關係可透過一條連續遞增的數學函數表示，稱為試題特徵曲線。每一試題都有各自的特徵曲線，因此試題的難易度與能力值則是在同一個量尺上的相對概念（郭伯臣、吳慧珉、陳俊華，2012；陳柏熹，2011）。CTT 假定所有受試者有相同的測量標準誤差，而 IRT 對不同能力的受試者提供不同的標準

誤差，因此能精確地推估受試者的能力估計值。CCT 對一份測驗的難度、鑑別度和信度等值，會受到不同受試樣本而改變；但是 IRT 的這些試題參數並不因樣本不同而有所影響。另外，依據 CTT，受試者在整份測驗上的總分即代表他的能力，而忽略受試者在試題的反應型態（item response pattern）。也就是得分相同的受試者，其試題反應型態可能不同，所反應出來的能力估計值也應有所不同。IRT 則可依據受試者試題反應型態的不同，而得到不同的能力估計值。對於一份測驗混和多元計分和二元計分的題型，CTT 對總分的計算會產生偏誤，而 IRT 卻能提供最佳的估計效果。再者，根據 CTT，受試者須透過常模參照的方式解釋分數在整體中的意義，但是，IRT 能將試題難度和受試者能力擺在同一個量尺，有助於受試者了解自己的能力和可能答對的題目（余民寧，2009；郭伯臣等人，2012）。

IRT 模式依照計分方式可分為二元（dichotomous）計分和多元（polytomous）計分模式。二元計分模式用於試題的答案反應僅有答對（1 分），答錯（0 分）兩種的測驗。多元計分模式是指題目的答題結果可能給予不同的分數，例如問答題可能有答錯（0 分）、部分答對（1 分）、全對（2 分）。二元計分模式可依據試題的參數個數分為單參數的 Rasch 模式、二參數模式和三參數模式。常見的多元計分模式有評定量尺模式（rating scale model, RSM）和部分給分模式（partial credit model, PCM）。評定量尺模式適用於量表中所有題目具有相同計分方式，如四點或五點 Likert 量表。部分給分模式是由 Masters（1982）提出，是 Rasch 模式多點計分的一個應用，適用於題目的評分點有次序概念的混合計分。

IRT 的發展常以試題特徵曲線中題目的參數個數，在分類上也常將 IRT 模式分為單參數模式、二參數模式和三參數模式，每個模式依序包含難度、鑑別度、猜對率等不同數量的參數。最早發展的單參數模式又稱為 Rasch 模式，以紀念創始者 George Rasch。在 Rasch 模式中，將題目的難度視為主要影響受測者答對機率的試題特性，要求所有試題需有足夠的鑑別度，且受測者猜對題目的機率被納入受測者的能力中考量，也就是其數學模式中僅包含受測者能力和試題難度單一個參數（陳柏熹，2011）。由於 Rasch 模式中題目的難度並不會受到受測者的能力水平而變化，具有明確的客觀性。王文中（1996）認為 Rasch 模式才是測量的模式，而二參數模式和三參數模式只是統計模式，沒有測量的意義。從 Rasch 模式發展後，越來越多學者延伸發展多元計分的各種 Rasch 測量模式，如 RSM、PCM，甚至多面向模式，以及適用於多向度的 Rasch 模型等，被稱為 Rasch 家族測量模式（余民寧，2009），可知 Rasch 測量模式在

測量和教育研究領域已受到廣泛地應用。

單向度（unidimensionality）是 IRT 的基本假設。單向度意指同一份測驗所有題目主要測量相同的某一項特質或構念。然而，為了因應同一份測驗中同時測量不同的特質或構念，例如智力測驗或綜合能力測驗（如自然科、社會科），測驗學者將 IRT 擴展為多向度（multidimentional）試題反應理論模式分析這類的測驗資料，例如現代公民核心素養測驗中包含倫理素養、科學素養、媒體素養、美學素養、民主素養等多向度素養的試題（陳彥君、洪素蘋，2019）。甚至，科學探究能力測驗中包含提問、實驗、分析和解釋四種能力的評量，採用四向度多元計分模式分析試題（Kuo et al., 2015）。單一向度檢定可做為決定採用單向度或多向度試題反映理論模式的依據。一般採用因素分析方法進行單一向度的考驗，並繪製特徵曲線圖做為判斷。評判方式有二，倘若第一因素的解釋變異量大於 20%以上，且第一因素解釋變異量明顯大於第二因素（Reckase, 1979）；或是第一因素特徵值大於第二因素特徵值的 2.5 倍時，且第二因素與其他因素差不多時（洪碧霞、吳裕益、林哲彥、葉千綺，1992），均可將此測驗資料評定為單向度。

Wright map 是運用 IRT 分析評量工具的另一個優點，其可將整份試卷的試題難易度圖像化，而非以表格方式呈現試題資訊。Wright map 或稱為人員－試題圖（person-item map），是以此圖的開發者 Ben Wright 命名，用來展現 Rasch 模式單向度或多向度評量結果（Wilson, 2005）。圖的左邊展現整體受試者的能力表現分布，圖的右邊呈現所有試題的難易程度。圖左越下方代表能力越低，越上方能力越高。圖右越下方代表試題越容易，越上方代表試題越困難，也就是將受試者在此評量的能力表現和試題的難易度共同呈現的一張圖。可做為評估試題難易度、建構反應試題配分的合適性、及試卷與學生能力的匹配度等方面之應用（Stelmack et al., 2004）。

參、研究設計

一、物理素養試題的發展與效化

以下將介紹物理學科中心種子教師所發展物理素養試題，經過選題、預試、專家審查、修改、施測等，再將施測結果進行 Rasch 分析，檢驗各題的適配性，以進行刪

題或修題，形成有效評量高三學生物理素養試卷的過程。

（一）試卷初稿的發展與效化

物理學科中心試題組的種子教師是來自全臺各縣市任教普通型高級中等學校班級的物理教師，大約 15 位，大多有任教高中物理多年的經驗。每個月試題組教師進行工作會議，種子教師每人發展至少一個題組的試題，第一作者藉此引導教師們腦力激盪，在題組情境下發展小題，或建議修改題目的意見。發展足夠的試題後，試題組團隊共同選出 15 個題組，42 小題的試題，並將其分為難易程度、試題數量相當的 A、B 兩試卷，施測時間均為 1 小時。物理素養試卷目的為評量高中學生在生活情境中，應用高一必修物理課程知識以解決問題的能力。種子教師依據表 1 的內容知識發想相關的情境與問題，通常發展解釋面向的題目大多運用內容知識，設計面向的題目則運用程序性知識，而詮釋面向的題目則會使用一部分的程序性知識，如「變因的概念」和「將數據抽象化或呈現的常用方法，如表格、圖形……。」，也可能應用內容知識。因此，表 2 利用 108 課綱中自然領綱高中物理必修課程的學習內容次主題為核心概念（如表 1），檢視兩試卷應用核心概念展現科學能力的題目數量。兩試卷涵蓋物理必修課程中八成以上的物理核心概念，並且分別應用不同的核心概念。

表 2　物理素養 A、B 試卷核心概念與科學能力的雙向細目表

	A 卷			B 卷		
物理核心概念	解釋	設計	詮釋	解釋	設計	詮釋
能量的形式與轉換	1		4			
溫度與熱量				1	2	1
自然界的尺度與單位	1		1			
力與運動				4	2	5
波動、光及聲音				1		
萬有引力	2					
電磁現象	1	1	2	2		1
量子現象	3					
科學在生活中的應用	5			2		
總題數	13	1	7	10	4	7

註：「解釋」、「設計」和「詮釋」三種能力為 PISA2015 科學素養評量架構中「解釋科學現象」、「評估及設計科學探究」和「詮釋科學數據與證據」的簡稱。

A、B 試卷初稿分別由來自南部兩個公立社區型高中 105 位和 106 位高三學生進行預試，由學生答題結果，初步檢驗題目的可讀性、試題的難易程度，然後修改選擇題的選項和建構試題的評分標準。另外，兩試卷初稿送給兩位物理專長的教授，以及另外一位熟悉 PISA 試題的物理教育教授審查。審查意見主要為題目的文字表達，力求精準。其中一位專家特別指出大多數試題為評量物理素養的試題，但是少數題目仍難以跳脫傳統評量學生運用公式即可解答，缺乏應用物理知識的設計。然而，在高中物理課程中，不少核心概念可用公式表徵變因之間的關係，例如手機以光速傳送電磁波，須利用光速與電磁波頻率的關係式（$C = \lambda \times f$）推論電磁波的波長。試題組團隊認為這些簡易公式所內含的核心概念屬於物理素養，因此，仍保留這些題目進行施測，後續將針對這兩類題目的施測結果進行討論。

試題組教師依據三位專家的審查意見再次修改試題與評分標準後，成為正式施測的試卷。A 卷共 21 題，包含 14 題單選題，1 題複選題，6 題建構反應題；B 卷共 21 題，包括 12 題單選題，1 題複選題，1 題是非題，7 題建構反應題。

以下舉 B 卷選擇題組「汽車的空氣阻力」中其中一題建構反應題，和 B 卷三個建構反應題構成的題組「汽車撞擊力廣告」，與其評分標準為範例。其中，「汽車的空氣阻力」的第 11 題用來評量學生運用物理概念設計降低汽車空氣阻力的兩種方法，屬於評量設計科學探究的能力。而「汽車撞擊力廣告」的第 17 題則係評量學生能否運用自由落體運動公式，以檢驗廣告中汽車著地速度，是評量解釋科學現象的能力。第 18 和 19 題則係評量學生將數據轉換成 X─Y 圖，以及解讀 X─Y 圖傳達物理意涵的能力，屬於評量詮釋數據與舉證的能力。

1. B 卷【汽車的空氣阻力】（原第 11 題，刪題後為第 10 題）

汽車行駛時，需克服地面摩擦阻力和空氣阻力，若要以較快車速行駛，則引擎要輸出的能量便要愈大，汽車也就愈耗油。一般汽車行駛時的所遇到地面摩擦阻力幾乎為定值，因此在某一速度以上阻力主要就來自空氣阻力。

空氣阻力來源大致分為三種形式：

一、正面阻力：氣流撞擊車輛正面所產生的阻力，此阻力來自於擋風玻璃、進氣格柵等部件。

二、表面摩擦阻力：當車輛以最快速度行駛時，表面摩擦阻力小到幾乎可以忽略。

三、外型阻力：車輛高速行駛時後部真空區形成的外型阻力，外型所造成的阻力來自車後方的真空區，真空區越大，阻力就越大；外型阻力是最主要的空氣阻力來源。

汽車車身的設計如果不流暢就會增加外型阻力；風阻係數是可根據車身的形狀改變而減小的。下表是已知形體的風阻係數：

物體或形狀	風阻係數
垂直平面體	1.0
球體	0.5
噴射飛機	0.08

空氣阻力示意圖

問題 11：現今太陽能動力賽車將太陽能轉化為電能而驅動車輛，除要減輕車重外並要能有較大受光面積，同時要能降低風阻係數，而經精密的設計，已有太陽能動力賽車風阻係數低到只有 0.07。請問，在設計太陽能動力賽車（如下圖）時，設計者如何改變車體來降低空氣阻力?請舉出兩種具體方法。

（太陽能動力賽車實體）

評分標準

滿分 2：

學生寫出「減少氣流撞擊車輛正面所產生阻力的方法」，可用來探討主要能降低風阻係數的關係。答案提及到「減少正面立面面積」、「流暢的車身設計」、「改變車輛形體」、「車尾立面面積」等，主要能降低風阻係數的因素兩項或以上。

部分得分 1：下列類型之一的正確答案

答案提及到「減少正面立面面積」、「流暢的車身設計」、「改變車輛形體」、「車尾立面面積」等，主要能降低風阻係數的因素其中一項。

未得分 0：錯誤、不明確或是無關答案。如

（1）學生直接陳述非相關因素，如「加大馬力、增加太陽能板」等。

（2）不符合太陽能車需有大面積受光面積的原則等答案。

2. B 卷【汽車撞擊力廣告】（原第 17~19 題，刪題後為第 15~17 題）

汽車廠商為強調安全性，將汽車吊離地面，再使汽車從高處落下，撞擊地面後，若仍可輕易開啟車門，表示車體剛性及對車內乘客在受到撞擊的保護能力較佳。假設車頭吊離地高 9.8 公尺，落下後撞擊地面，車商表示：『撞擊力相當於時速 50 公里。』

問題 17：車頭垂直離地高 9.8 公尺，車子加速度為 9.8（m/s^2），車頭落地瞬間速率為多少（km/h）？需列式計算，$\sqrt{2}=1.41$

評分標準

滿分 2：學生正確寫出計算過程和答案者，

$$v^2=0^2+2\times9.8\times9.8\Rightarrow V=13.86(m/s)=49.86(km/h)$$

(1)　　(2)　　(3)

部分得分 1：

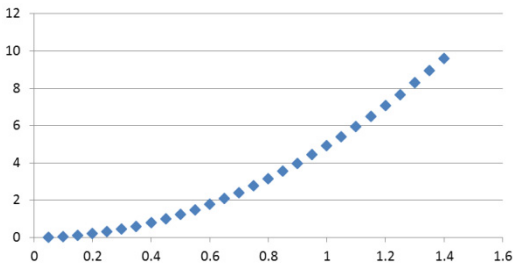
答出上方答案三部分的其中之一，得到部分分數一分，如答案正確（13.86m/s 或 49.86km/h 或接近的答案），或僅列式正確。

未得分 0：錯誤或無關答案。

問題 18：某生測量此撞擊測試影片的數據，並做出圖一之結果，請問圖一應是哪二種物理量之關係？

橫軸物理量：＿＿＿＿＿

縱軸物理量：＿＿＿＿＿

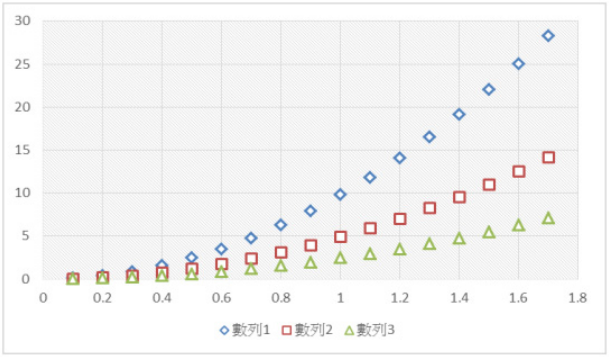


圖一

評分標準

滿分 2：縱軸是位移，橫軸時間，兩個答案都正確才給分。
部分得分 1：兩個答案答對其中一個。
未得分 0：錯誤答案或未答者。

問題 19：圖二之橫軸與縱軸物理量與圖一相同，今將一顆 20kg 大鐵球及一顆 5kg 小鐵球自 9.8 公尺自由落下，其物理量表示為數列 1、數列 2、數列 3 之哪個數列？已知汽車落下物理量關係為數列 2。



圖二

大鐵球：□數列 1,□數列 2,□數列 3
小鐵球：□數列 1,□數列 2,□數列 3
理由：_____

評分標準

滿分 2：兩球可視為自由落體，故皆為數列 2。答案和理由都正確才給分。
部分得分 1：僅答案正確（皆為數列 2）。
未得分 0：答案錯誤或未填答。

（二）試卷施測與資料分析

施測樣本乃依據全臺灣普通型高中入學分數百分等級（Percentile Rank, PR）進行分層抽樣，將學校分高、中、低 PR 三群，並以方便取樣，在物理學科中心試題組教師所任教的學校中挑選代表三群 PR 的學校。最後共 19 所學校，35 班，1119 位高三學生參與施測。經刪除無效試卷後，554 份 A 卷和 560 份 B 卷進行分析。參與 A 卷高、

中、低 PR 組學生數分別為 130、328、96 人，參與 B 卷高、中、低 PR 組學生數分別為 127、299、134 人。

由於物理素養 A、B 試卷中，部分開放性問題為多元計分試題，因此，將學生答題的數據運用 ConQuest 軟體的部分計分模式（partial credit models, PCM）進行試題分析。每一題建構試題都有兩位教師依據評分標準評分，評分有差異的答案則交由第三位教師評分，以決定最終得分。A、B 試卷的有效資料先經過因素分析，進行單向度檢驗，以決定採用單向度 PCM 或多向度 PCM。以第一因素的解釋變異量是否大於 20%，以及第一因素特徵值第二因素特徵值是否大於 2.5 倍為標準。若否定假設，則採用多向度 PCM（洪碧霞等人，1992；Reckase, 1979）。

評量工具所關切的「變項」引導著工具發展者發展出與變項有關的題目，這些「變項」就是某種理論構念／建構（construct）。王文中（1997）認為 Rasch 分析最有價值的功能就是利用題目和受試者的反應，驗證這些題目能否有效地測量某種理論構念／建構，如物理素養、科學態度等。倘若實證資料吻合 Rasch 模式，這些題目便具有測量構念／建構的功能。本研究利用專家審查和預試分析確保 A、B 兩試卷的內容效度，及每一小題評量某種科學能力的代表性（林小慧、林世華、吳心楷，2018）。另外，為了評估兩試卷結構上的建構效度，須檢視各面向試題的難度、Wright map 上各面向難度分配的廣度和受試者能力的對應狀況，以及預設作答表現的層級和對應試題難度間的一致性（林小慧等人；Kuo et al., 2015）。

適配度分析（goodness-of-fit analysis）是 Rasch 家族測驗模型運用來檢定資料是否適配於所使用的測量模型，主要以均方（mean square, MNSQ）值介於 0.7─1.3 和 *t* 分配指標的絕對值小於 2 為評判試題是否符合模型的基準。超過此範圍的試題表示此試題已達不適配程度，需經過刪題或修改評分標準再重新分析，以獲得適配度令人滿意的試題分析結果。另外，Infit MNSQ 值為加權的均方，對於受試者與試題互相匹配具有敏感性，在偵測不適配的試題上，Infit MNSQ 值會比 Outfit MNSQ 值受到更多的重視（余民寧，2020；Wu & Adams, 2007）。Rasch 模式常以估算快速且誤差較小的貝氏期望後驗來估計（Expected A Posteriori（EAP）estimate）受試者的能力，而以 EAP 估計結果作為受試者的能力參數估計值（單位 logits）。因此，兩試卷的信度值將採用 Expected A Posteriori/Plausible Value （EAP/PV）reliability。再者，試題分離信度（separation reliability）也是 Rasch 模式用來了解測驗結果好壞的指標，分離信度越高，試題難度差異越明顯，越容易分辨出受試者能力。

肆、研究結果

一、建構效度評估結果

A 卷資料經因素分析後，第一因素的變異數解釋量為 11.5%，未達 20%，而且第一因素與第二因素的特徵值比值（3.15／1.53= 2.06）未達 2.5，不符合單一向度，將依據發展 A 卷的三面向評量架構，評估 A 卷的測驗模型。然而，A 卷中評量設計科學探究能力的題目僅有一題，不適合單獨成為一個向度。因此，將設計與詮釋面向的題目結合為一個向度，進行二向度 PCM。表 3 和表 4 是將學生的應答結果經多向度 PCM 運算，而得到的分析結果。表 3 顯示 A 卷幾乎所有試題的難度參數估計值介於 -1.5 至 1.2 之間，而且每一試題的 Infit MNSQ 值均在 1$\pm$0.3 範圍之間，且其 *t* 值亦都小於 2，顯示 A 卷已達到適配程度。分離信度為 .995，顯示試題難易差異明顯。唯獨第 17o2 題的難度估計值（2.285）超過 1.5，而且在表 4 第 17 題第二步驟 Outfit *t* 值 4.1，超過 2。此題目為「藍光與眼睛」題組中一題建構反應題（題號 17o2），原本訂定滿分為 2 分，需寫出兩個可能造成視網膜病變的原因，如「藍光能量較紅光強」和「直視強光源照到視網膜之輻射」。顯然獲得滿分兩分的困難度頗高，因此，此題評分標準的滿分修改為 1 分，寫出其中一個正確因素即可得滿分。此題修改後題號改為 17o（表 3）。

表 3　物理素養 A 卷二向度多元計分模式的試題難度參數估計值（*N*=554）

題組名稱	題號（修改後）	面向	難度	Outfit		Infit		備註
				MNSQ	*t*	MNSQ	*t*	
電影中的物理	1	解釋	0.405	1.00	-0.0	0.99	-0.4	
	2	解釋	1.132	1.02	0.4	1.01	0.6	
斷電系統	3	解釋	-0.176	0.94	-0.9	0.99	-0.2	
	4	解釋	0.082	0.92	-1.4	0.96	-1.1	
	5	解釋	0.286	1.02	0.3	1.00	0.1	
	6	解釋	-1.485	0.79	-3.7	0.95	-0.4	
盪鞦韆	7	詮釋	-1.432	1.05	0.8	0.99	-0.1	
	8	詮釋	0.204	0.98	-0.3	1.01	0.2	
電器廣告	9	解釋	-0.127	0.97	-0.5	0.99	-0.1	
光微處理器	10	詮釋	0.316	0.99	-0.1	0.99	-0.3	
	11m2	解釋	0.072	1.05	0.8	1.05	0.8	
運動飲料	12o	詮釋	-0.346	0.86	-2.4	0.91	-1.7	
	13o2	詮釋	-0.901	1.15	2.4	1.08	0.9	
	14o2	詮釋	0.676	0.91	-1.5	0.98	-0.4	
藍光與眼睛	15o2	解釋	0.583	1.10	1.6	1.06	1.4	
	16	解釋	-1.552	0.88	-2.1	0.95	-0.4	
	17o2（17o）	解釋	2.285	1.07	1.1	1.07	1.3	修改評分
電流磁效應	18	解釋	-1.506	0.88	-2.0	0.99	-0.1	
	19	設計	1.125	1.00	0.0	0.99	-0.5	
	20	詮釋	-0.405	0.87	-2.3	0.92	-1.6	
	21o2	詮釋	0.764	1.11	1.7	1.08	1.7	

註：試題題號後方標示英文字母 m 代表複選題，o 代表建構反應題，若英文字母後方附加 2 代表該題最高得分 2 分，部分得分 1 分。若未標示英文字母者為單選題，未附加 2 代表該題得分 1 分。

表 4　物理素養 A 卷二向度多元計分模式的試題步驟參數估計值（*N*=554）

題組名稱	題號_step	難度	Outfit		Infit		備註
			MNSQ	*t*	MNSQ	*t*	
光微處理器	11o2_0		0.87	-2.2	1.00	0.0	
	11o2_1	-1.748	1.05	0.8	1.04	1.4	
	11o2_2	1.748	1.08	1.4	1.06	1.4	
運動飲料	13m2_0		1.13	2.2	1.06	0.4	
	13m2_1	0.066	1.06	1.0	1.01	0.1	
	13m2_2	-0.066	1.10	1.6	1.05	0.8	
	14o2_0		0.90	-1.6	0.99	-0.3	
	14o2_1	1.129	0.97	-0.4	0.99	-0.0	
	14o2_2	-1.129	0.93	-1.2	0.98	-0.4	
藍光與眼睛	15o2_0		1.25	3.9	1.07	1.3	
	15o2_1	-0.259	0.98	-0.4	0.98	-0.6	
	15o2_2	0.259	1.02	0.4	1.01	0.5	
	17o2_0		1.09	1.5	1.05	1.8	修改評分
	17o2_1	-1.747	1.04	0.7	1.04	1.6	
	17o2_2	1.747	1.26	4.1	1.04	0.3	
電流磁效應	21o2_0		1.11	1.8	1.07	1.5	
	21o2_1	0.002	1.01	0.2	1.01	0.2	
	21o2_2	-0.002	1.13	2.1	1.05	1.4	

B 卷資料經因素分析後，第一因素的解釋變異量為 12.30%，第一因素與第二因素的特徵值比值（3.29／1.47=2.24）未大於 2.5，未通過單一向度檢定。因此，採用三向度 PCM 評估 B 卷的測驗模型。表 5 和表 6 為 B 卷經三向度多元計分模式分析得到的難度參數估計值。除了第 1y、4、5、10 題外，大多數試題的難度值域在±1.5 之間。另外，除了第 10 題和第 17o2 題的 Infit *t* 值和 Outfit *t* 值絕對值均大於 2 以外，所有試題的 Infit MNSQ 均在 1±0.3 範圍內，顯示僅有少數試題需要刪除或修改，絕大多數的試題是適配的。B 卷的分離信度為 .997，顯示試題難易明顯。第 10 題有關「虹與霓」的選擇題，學生須看得懂解釋虹與霓現象的光線分析圖，對大多數學生頗為困難（難

度值 1.47），決定刪除此題。從前述的試題可看到第 17 題為利用自由落體公式檢驗汽車著地的速度，若正確寫出計算過程與答案者得滿分 2 分，僅列公式或答案正確可得部分分數 1 分。最後修改此題評分標準，正確寫出計算過程與答案者得滿分 1 分，其餘答案為 0 分。

表 5　物理素養 B 卷三向度多元計分模式的試題難度參數估計值（*N*=560）

題組名稱	題號（修改後）	面向	難度	Outfit		Infit		備註
				MNSQ	*t*	MNSQ	*t*	
保溫瓶	1y	解釋	-1.970	1.13	2.2	1.04	0.3	
	2	設計	0.962	1.07	1.1	1.03	1.0	
	3	詮釋	-0.601	1.07	1.1	1.03	0.6	
	4	設計	-1.982	0.70	-5.7	0.88	-0.6	
行車安全	5	詮釋	-1.564	0.93	-1.2	0.96	-0.5	
	6	解釋	-1.053	0.87	-2.3	0.93	-0.9	
	7	詮釋	-0.861	1.03	0.4	1.04	0.7	
手機電磁波	8	解釋	0.170	0.95	-0.9	0.97	-0.9	
	9	詮釋	0.924	0.91	-1.5	0.94	-1.8	
虹與霓	10	解釋	1.514	1.20	3.2	1.16	4.0	刪題
汽車的空氣阻力	11o2（10o2）	設計	1.521	1.04	0.6	1.04	0.7	
	12（11）	詮釋	-0.508	1.02	0.4	1.01	0.2	
	13m2（12m2）	設計	-0.501	1.00	0.0	0.98	-0.3	
	14（13）	詮釋	0.690	1.01	0.2	1.00	-0.1	
口罩	15o（14o）	解釋	-0.518	0.83	-3.0	0.91	-1.8	
	16o	解釋	0.222	1.07	1.1	1.06	1.8	刪題
汽車撞擊力	17o2（15o）	解釋	0.612	0.85	-2.6	0.88	-2.8	修改評分
廣告	18o2（16o2）	詮釋	1.210	1.07	1.2	1.03	0.5	
	19o2（17o2）	詮釋	0.710	0.99	-0.2	0.97	-0.7	
馬桶	20（18）	解釋	-0.008	1.02	0.4	1.01	0.4	
	21o（19o）	解釋	1.030	1.04	0.7	1.04	1.3	

註 1：試題題號後方標示英文字母 y 代表是非題，m 代表複選題，o 代表建構反應題，若英文字母後方附加 2 代表該題最高得分 2 分，部分得分 1 分。若未標示英文字母者為單選題，未附加 2 代表該題得分 1 分。

註 2：此試題在刪除第 10 題和修改第 17 題評分標準後，經多向度 PCM 分析結果 Outfit *t* 值和 Infit *t* 值均大於 2 而刪題。

表 6　物理素養 B 卷三向度多元計分模式的試題步驟參數估計值（*N*=560）

題組名稱	題號_step	難度	Outfit		Infit		備註
			MNSQ	*t*	MNSQ	*t*	
汽車風阻	11o2_0		1.03	0.6	0.99	-0.2	
	11o2_1	-1.559	1.01	0.1	1.01	0.2	
	11o2_2	1.559	1.11	1.7	1.07	0.9	
	13m2_0		1.23	3.6	0.91	-0.3	
	13m2_1	-1.235	1.00	0.1	1.00	0.1	
	13m2_2	1.235	1.00	0.1	1.00	0.1	
汽車撞擊力廣告	17o2_0		0.77	-4.2	0.87	-3.2	
	17o2_1	0.111	0.96	-0.7	0.98	-0.5	
	17o2_2	-0.111	0.94	-1.0	0.94	-1.5	
	18o2_0		1.06	1.0	1.01	0.3	
	18o2_1	1.713	1.03	0.5	1.00	0.1	
	18o2_2	-1.713	1.08	1.3	1.05	0.8	
	19o2_0		0.99	-0.2	0.98	-0.4	
	19o2_1	1.473	1.03	0.4	1.00	0.1	
	19o2_2	-1.473	0.99	-0.1	0.96	-1.0	

刪除第 10 題和修改第 10 題評分資料，重新執行三向度 PCM 後，在 ConQuest 試題報表中顯示第 16o 題的 Infit *t* 值和 Outfit *t* 值絕對值均大於 2，因此，再度刪除此試題，獲得最後版的 B 卷。

經過試題資料分析後，A 卷修改 1 題評分標準，B 卷刪除 2 題，修改 1 題評分標準。最後版的 A 卷共 21 題，B 卷共 19 題，詳細的題型數量，以及評量三種科學能力的題目數量如表 7 所示。

表 7 呈現兩份物理素養試卷效化後的題目數量和試卷難度、信度值與 MNSQ 值。很明顯地，兩試卷中評量解釋科學現象和詮釋數據與舉證的試題較多，而評量評估與設計科學探究能力的試題較少。B 卷設計向度題目較少（僅 4 題），可能也導致此向度評估受試者科學能力 EAP/PV 信度值偏低。由於 A、B 兩試卷的 Cronbach's α 分別為 .69 和 .70，達評量工具可接受的最低限度（DeVellis, 2017; Nunnally, 1978）。然而，

兩試卷的分離信度頗高（A 卷 .993，B 卷 .997），難度與 Infit MNSQ 值頗佳，顯示兩試卷大致上可評量出學生不同的物理能力。

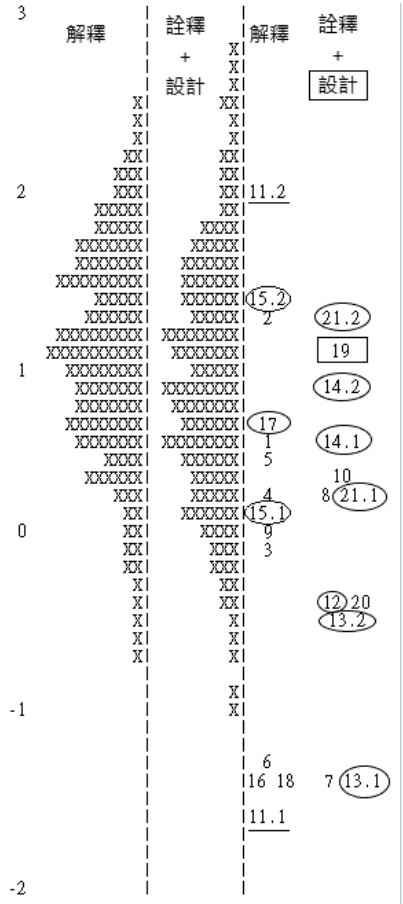
表 7　效化後物理素養 A、B 試卷的試題類型數量與試卷難度、信度值與 MNSQ 值

	A 卷			B 卷		
題數[註]	21			19		
總分	26 分			23 分		
分離信度	.993			.997		
Cronbach's α	.69			.70		
能力面向	解釋	設計	詮釋	解釋	設計	詮釋
題數	12	1	8	7	4	8
EAP/PV 信度值	.653	.688		.669	.533	.651
難度平均值（範圍）	-0.09（-1.41−1.29）	0.12（-1.43−1.23）		-0.05（-1.89−1.42）	0（-1.99−1.52）	0.12（-1.56−1.21）
Infit MNSQ 平均值（範圍）	1.03（0.95−1.06）	0.99（0.91−1.09）		1.01（0.93−1.08）	0.98（0.90−1.05）	1.00（0.96−1.03）

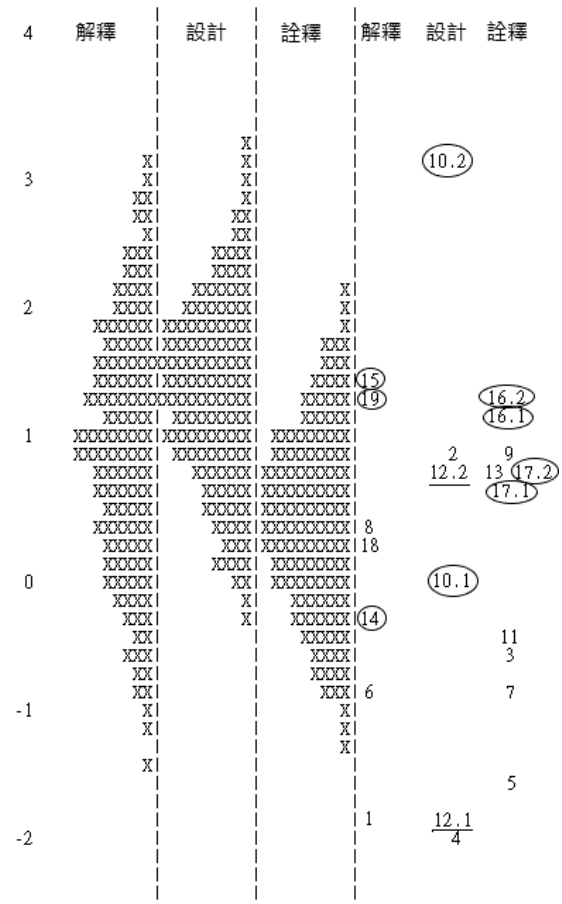
註：A 卷包含 14 單選題，6 問答題，1 複選題；B 卷包含 11 單選題，6 問答題，1 複選題，1 是非題。

二、Wright map 在物理素養試題分析上的應用

Wright map 可呈現 Rasch 多向度部分給分模式整體受試者在各向度所呈現的能力分布，以及試題難易度分布。兩試卷分別由試題數量 30 倍以上具有常模代表性的受試者所施測。圖 2（A）和（B）為刪除不合適試題後，A 卷和 B 卷受試學生得分表現的 Wright map。



（A）每個 X 代表 4.0 位個案，方框為設計面向的題目



（B）每個 X 代表 4.2 位個案

圖 2 （A）A 卷和（B）B 卷受試者能力與各面向試題難度的（步驟參數）Wright map

註：圓圈的題目為建構反應題，畫底線的題目為複選題，未標示者為單選題。

圖 2（A）顯示學生在 A 卷兩個向度能力表現呈現常態分布，平均能力大致在 1.0（logit）附近（解釋：1.09，詮釋+設計：0.92），試題的難度平均值大致在 0 附近（解釋：-0.09，詮釋+設計：0.12），表示學生能力在這兩個向度表現相似，兩向度的試題難易適中。解釋向度的試題有 12 題，試題分布顯示都有高、中、低難度的試題，大

致可評量到不同能力的學生。相較之下，「詮釋+設計」向度僅有 9 題，且較少高難度的試題能合適地評量高能力的學生。但是，「詮釋+設計」向度 EAP/PV 信度值（.688）較解釋向度（.653）高，應該是「詮釋+設計」向度三層級評分的建構反應試題較多所致（解釋：0 題，詮釋+設計：3 題）。

圖 2（B）也顯示出學生在 B 卷三個向度能力表現均呈現常態分佈，在解釋和設計向度的平均能力分別為 0.97 和 1.35，在詮釋向度的平均能力較低，為 0.41（表 8）。三個向度的試題難度平均值都在 0 附近（解釋：-0.05，設計：0，詮釋：0.12），顯示三向度的試題難易適中。由於 B 卷的建構反應試題在各向度都佔有總題數的 1／4 以上，有助於區別出不同程度能力的學生。然而 B 卷高難度的題目數量也顯示較少，可能較難有效地評量高能力學生的表現。

表 8　A、B 兩卷受試者的平均分數、得分率，及在不同面向上的平均能力

	A 卷（*N*=554）		B 卷（*N*=560）		
滿分	26		23		
Mean（*SD*）	17.77（4.08）		13.88（3.87）		
平均得分率	68%		60%		
平均能力（*SD*）（*logit*）	1.02（0.57）		0.31（0.59）		
面向	解釋	設計+詮釋	解釋	設計	詮釋
平均能力（*SD*）（*logit*）	1.09（0.54）	0.92（0.63）	0.97（0.85）	1.35（0.57）	0.41（0.61）

另外，圖 2 中畫圈的題目顯示建構反應試題，從圖中顯示不論是兩層級計分或三層級計分，大致上屬於偏中難度以上試題，較容易作為區別大多數學生物理能力的題目，未來在素養導向試題應多設計這類題目。

Wright map 以圖像呈現各試題難度值的方式，更有利於試題發展者一覽整份試卷中試題難易度的分布情況。當初種子教師依據教學經驗推測優質試題的難易度，將其平均分配至兩份試卷，期望兩份試卷的難易度相當。從圖 2 顯示兩份試卷的試題大致均勻分布於難度-1.5 至 2.5 之間，且兩試卷各向度難度平均值均位於 0 附近，顯示兩試卷整體的難易程度相當，如教師所預期。但是，Wright map 中還呈現更多資訊，值

得試題發展者多理解，以增進發展素養導向試題的能力。以前述介紹的四題建構反應題為例，這四題雖然都屬於建構反應題，但是滿分和部分得分的難度值有十分相近和有所懸殊兩大類。圖 2（Ｂ）右側的第 16 題（16.1 和 16.2）第 17 題（17.1 和 17.2）兩題屬於滿分和部分得分的難度值十分相近的題目。第 17.1 和 17.2 為原來 B 卷第 19 題部分得分和滿分的難度值，前者評量能否將大、小鐵球與汽車視為自由落體的概念，而滿分則是需要敘述正確理由。17.1 和 17.2 的難度值十分接近，顯示滿分和未得分的人數占大多數，部分得分的人數較少。也就是說，倘若學生能選對大、小鐵球在圖上的曲線，絕大多數也知道是運用自由落體的概念。同樣地，圖 2（Ｂ）右側的第 16 題為原來 B 卷第 18 題，學生若能答對 X 軸座標，大概也能正確回答 Y 軸座標，所以部分得分者較少。值得教師注意的是，這兩題均為評量學生將實驗數據轉換至 X─Y 軸圖的能力，及理解 X─Y 軸圖的意義。沒想到確認 X─Y 軸圖座標軸的名稱對大多數學生的難度頗高（僅 22%得滿分，72%零分），幾乎與第 15 題計算自由落體速度的題目難度值相當。這資料也顯示高中生在學習階段可能很少有機會將數據繪製作圖，以致無法正確寫出座標軸的物理量。

圖 2（Ｂ）右側的第 10.1 和 10.2 為原來 B 卷第 11 題部分得分和滿分的難度值，兩個難度值相差 3（logit）。比較原題目第 11 題與第 18、19 題的評分標準，發現第 11 題需要有較多文字的表達，闡述出至少兩種降低空氣阻力的方法。正確答案須能展現出運用至少兩種主要概念。由這兩類建構反應題可知，使用較多文字回答問題的題目，較能區別出學生科學溝通的表現，因此部分得分與滿分的難度值差異較大，較能有效評量出高、低不同程度學生的能力。相反地，類似簡答題的建構反應題在評量不同程度學生的效力上可能較差。除了建構反應試題外，在圖 2 中劃底線的複選題，包括 A 卷修改後的第 11 題和 B 卷修改後的第 12 題，兩個階層的難度值相差大約 2（logit）以上，也能有效評量不同能力的學生。優質的評量工具在 Wright map 上應該是右側各種難度試題的分布能對應涵蓋左側各種不同能力的學生。因此，從圖 2 A、B 卷的 Wright map 顯示，兩卷若再增加數題難度值較高的題目，會成為更好的評量工具。

伍、結語

物理素養評量主要目的為檢視高中生應用物理知識，展現其解決問題的能力。本

研究利用 PISA2015 評量科學素養架構（OECD, 2017），發展能有效評量學生物理素養的試題。利用 Rasch 多向度部分給分模式，分析出試題組團隊所發展的試題大多數具有良好的配適度，難度適中，尚可有效評量高中生的物理素養。不過，本研究在研究設計與施測上尚有不夠嚴謹之處，造成一些研究限制。另外，在發展與效化試卷的過程和參考相關文獻中，也發現一些值得關注的重要議題，共列出五點，可做為未來教育研究者推動教師發展素養試題之參考。

一、孰為合適的物理素養評量架構？

本研究所介紹的物理素養試題乃自 105 學年度物理學科中心試題組團隊開始發展的，當時尚未頒布 108 自然領綱，因此採用 PISA 科學素養評量架構，作為發展能夠評量學生在各種情境中應用物理內容知識和程序性知識，展現解釋科學現象、評估與設計科學探究，及詮釋數據與科學舉證的三大能力。林蓓伶等人（2018）依據 108 自然領綱，將其中的「學習表現」作為設計科學素養題型的目標，將「學習內容」作為試題的主題範圍，主要評量六項「探究能力」（即審視資訊並界定問題、規劃實驗／探究計畫、執行實驗／探究計畫、分析與陳述數據、解釋數據並說明推理過程、分享與評鑑實驗／探究結論），或五項「科學本質與態度」（覺察、動機與興趣、審視與評價、參與科學議題、建立科學自我效能感）作為科學素養評量架構。盧政良、林蓓伶與何興中（2019）依據 108 自然領綱中探究與實作學習內容的四個面向（發現問題、規劃與研究、論證與建模、表達與分享），為發展高中物理探究與實作素養試題的架構。Adeleke 與 Joshua（2015）依據 Akindehin 所提出五大面向的科學素養，發展出高中物理素養試題，其五大面向包括自然現象的知識、物理基本概念、科學本質知識等三類知識，以及科學心智習慣／能力、解決問題與執行探究的能力等兩類能力。這四個研究中，盧政良等人和林蓓伶等人都是以 108 領綱「學習表現」作為素養評量之架構，著重於探究能力，而缺乏應用物理知識解釋現象的能力。從國內外對科學素養的定義與內涵中都曾表示應用和表達科學知識乃為科學素養的主要表現之一（教育部，2018；Bybee, 1997；NRC, 1996；OECD, 2017）。再者，PISA2006 和 PISA2015 科學素養評量中，評量解釋科學現象的題目佔總題數的五成，設計和詮釋面向的題目約各佔兩成和三成（佘曉清、黃莉郁、蘇怡蓓，2017；Bybee et al., 2009）。解釋面向的題目大多以評量應用「內容知識」的能力，設計和詮釋面向的題目則大多數評量應用程序性知識和認識論知識的能力。內容知識即是科學核心概念，也就是 108 自然領綱中

所謂的「學習內容」。「學習表現」包含探究能力與科學的態度與本質。由於科學本質包含程序性知識、認識論知識及科學本質觀，前面兩種知識可展現在探究能力上，於 PISA 的設計和詮釋面向的試題所測得。科學本質觀和科學態度屬於情意面向，有合適的「科學本質觀」和「科學態度」工具評量之。因此，林蓓伶等人將「科學本質與態度」納入科學素養評量架構中，將情意面向的得分和探究能力混合計分是否恰當，應慎重再考量。總之，108 自然領綱兼重「學習內容」和「學習表現」兩部分，與 PISA 科學素養評量架構的內涵是一致的。目前大學入學考試中心（2020）所公告素養導向命題方向因考量所有領域的特質，情境題的設計偏重於閱讀素養所重視的能力（如擷取訊息、廣泛理解……等），未必需要應用科學知識所展現的科學能力。這恐怕會讓發展科學素養試題的教師，誤以為採用科學情境素材的題目即為評量科學素養的試題。108 課綱強調跨科、跨領域的能力，未來自然科學領域的素養試題，包含評量閱讀素養與科學素養等多元化能力，是否就是我國入學考試未來的方向？有待討論。108 課綱中各領域有其特定培養的能力，各學科中心是否應先凝聚共識，建立同一領域的素養評量架構？這不是推動素養評量的首要之務嗎？

二、運用重要的物理公式解釋科學現象可視為物理素養的展現？

在 A、B 卷物理素養試卷中有五題涉及物理公式，例如建構反應題示例「汽車撞擊力」第 17 題。對大多數物理教師而言，許多重要的公式相當於物理概念，屬於一種內容知識。因此，具備物理素養的學生，應可利用簡單的公式檢驗電視廣告的宣傳詞是否正確。然而，有些素養評量專家卻反對將套用公式便可解題的試題視為素養題目，例如一位審查試題的專家指出「偏向記憶性，套公式便可解題的題目」，不算是素養題目。他的另一個理由是，若因計算問題導致科學現象解釋錯誤，這類試題的效度便可能受到質疑。審查委員的想法主要受 PISA 科學素養評量所影響。PISA 施測對象是 15 歲學生，國三學生的科學素養適合運用中學的科學知識展現能力。然而，物理素養試卷係評量高三學生，數學運算是解決生活情境中的物理問題，應屬於基本能力，如同 PISA 數學素養也需要運用數學核心概念的公式展現數學能力。因此，建議這類因套用公式，需經過計算展現科學能力的試題，如同示例「汽車撞擊力」第 17 題，建立三個層級的評量標準，計算過程中運用公式表達理解核心概念者部分得分，正確計算獲得解答者得滿分，以符合高中學生應達到的物理素養。

三、發展建構反應試題、設計面向與高難度素養試題的重要性

過去研究所發展的物理素養評量工具都使用單選題一種題型（盧政良等人，2019；Adeleke & Joshua, 2015）。然而，一些學者建議評量工具中結合選擇題和建構反應試題，較能全面地評量到學生不同程度的能力（McCoubrie, 2004）。本研究分析兩份物理素養試卷的結果，也發現不少建構反應試題因具有三層級的評分標準，達到滿分的難度值大多位居中、高難度，如此，能有效評量中、高能力的學生。不過，從種子教師發展物理素養試題的經驗中發現，大多數教師不太擅長掌握開放答案的核心概念，以訂定建構反應試題的評分標準。未來相關單位舉辦教師發展素養導向試題的研習課程中，可多參考 OECD 所釋放出來的建構反應試題與評分標準發展試題，以及批改學生回答情形，與評分人員共同修改、精緻化評分標準等歷程，將可增進教師這部分的能力。

一般而言，教師最擅長發展解釋面向的題目，其次是詮釋面向，而設計面向的題目最為缺乏。此現象不僅呈現在本研究的兩份試卷上（如表 2），在 PISA2006 和 2015 評量科學素養的試題數量也有類似的結果（佘曉清、黃莉郁、蘇怡蓓，2017; Bybee et al., 2009）。研究者輔導種子教師們發展素養試題的歷程發現，高中物理教師長期以來深受大學入學考試所影響，習慣發展評量概念理解或概念應用於解題的兩大類試題。PISA 試題中，應用內容知識解釋科學現象的題目最相似於這類型的試題，所以試題數量最多。而詮釋面向的題目能從實驗數據或日常生活的資訊發展相關題目，大多數物理教師也重視詮釋數據的能力。然而，高中教師甚少提供學生探究取向活動的經驗，對教師而言，發展設計面向的試題較為困難。因此，物理素養試題庫中，此面向的題目極為缺乏。另外，種子教師從素養試題庫選擇優質試題後，主要以兩份試卷難易度相似分配題組，而未以各面向題目數量相當為主要考量。這部分的缺憾，值得未來試題發展者運用雙向細目表，考量試題類型的均勻分布，作為改進的參考。未來建議教師可從科學探究與實作活動中取材，從實作活動中重要的知識與能力發展試題。尤其，教師若能在教學中以素養導向的教學方式，提供學生表現探究能力的機會，並建立實作評量的評分標準，應有助於發展評量設計面向的試題（林春煌，2019）。

雖然本研究評量高三學生物理素養的兩份試卷平均難易度適中，但是在圖二 Wright map 顯示較少高難度值的試題。盧政良等人（2019）發展六個題組 16 題物理探究與實作素養試題，經 CTT 得到試題的平均難度值為 0.47，也是屬於難易適中。

而林蓓伶等人（2018）發展四個題組 14 題科學素養試題，利用 Rasch 模式得到的難度平均值為-1.32，絕大多數題目偏簡單。不少教師認為素養試題偏重概念應用，缺乏計算，試題難度較低。然而，從 OECD 所釋出的科學素養試題中，仍發現有評量學生分析、評估等高階思考能力的高難度題目。總之，建構反應試題、設計面向和高難度素養試題都是未來發展素養試題值得努力的方向。

四、善用 Wright map 增強教師對素養試題難度值與學生能力的了解

Wright map 能完整地呈現受試者能力和試題難易度分布，試題發展者能累積判讀每一個試題在此圖難度值的經驗，有助於增加教師對學生能力表現的預測力、對試題難易度的敏感性，以及提升修改素養試題的能力。因此，建議發展素養導向試題時，應尋求熟悉 IRT 的研究人員，解析施測試題的結果，增進試題發展教師對素養試題的敏感度。另外，研究者發現教師為了鼓勵學生回答開放性問題，會有過度給分的狀況，例如前述 B 卷第 17 題列式、答案正確或接近均給予部分分數的狀況。然而，此題經 Rasch 模式分析結果，適合改以兩層次的評分，較能合適地評估學生的能力。這部分需由研究人員判讀試題難度參數、步驟難度參數或類別特徵曲線圖等數據，以協助教師理解。甚至學生表達不清楚的答案，教師會自行推測學生的想法而給分，而有高估學生能力的現象。建議未來施測素養導向試卷後，可經熟悉 IRT 的研究人員分析資料，解析 Wright map 的結果，依據實證資料提供試題發展者調整評分標準的層級。

五、本研究的限制及對多數量試題施測的建議

為了讓學生在 60 分鐘內完成物理素養試卷的施測，本次從試題庫中選出優良的試題分配至兩本試卷，卻忽略兩份試卷應該要有重複的試題，以便不同試卷的重複試題可做為量測試題難易度和學生能力的基準，以整合所有試題的難易程度在一張 Wright map 上。而且，A、B 兩卷的施測對象僅考量各別地包含高、中、低 PR 值學校的學生。若能嚴謹地設計各校選取程度相當的兩班學生分別施測 A、B 兩卷，便可假設學生能力相似，試題指標較得以互相參照與比較。這兩個設計上的缺失為本研究的限制。另外，這類全國性的施測在實施上有些困難，平衡不完全區組設計（Balanced Incomplete Block design, BIB）或許可以解決一些問題。

平衡不完全區組設計是一種用於解決有限的作答時間、試題數量較多、又能避免受試者因作答到後面容易分心等特徵，影響作答表現的研究設計（Gonzalez &

Rutkowski, 2010; Kuo, et al., 2015）。若以本研究的 A、B 兩卷總題數 42 題為例，如表 9 所示，應先將試題區分為 A, B, C, D 四個區組（block），每個區組題目數 10—11 題，以組合成A＋B，B＋C，C＋D，和D＋A四本試卷。四組受試者包含高、中、低成就受試者，組成四組常態樣本，各接受一本試卷的施測。最後的施測結果可繪製為一張 Wright map，整體評估所有題目的品質。若欲比較不同年度學生素養表現的差異，亦可採用 BIB 設計，選擇 1—2 個區組的舊試題重複施測。期待本研究未盡完善的設計，經事後文獻檢討，可提供未來施測素養導向試卷之參考。

表 9　改進物理素養試卷的施測設計

受試者	試卷	區組
第一組	1	A＋B
第二組	2	B＋C
第三組	3	C＋D
第四組	4	D＋A

誌　謝

感謝匿名審查委員和編輯委員會提供寶貴修改意見，提升本文的品質。本研究為物理學科中心試題組種子教師與行政人員等協力完成發送試卷、施測、閱卷、登錄資料等工作，臺中市立豐原高中柯閔耀老師、桃園市立內壢高中劉詠薇老師共同發展試題、其他共同討論和閱卷的種子老師們，以及物理學科中心蔡沛霖先生等行政人員的協助，特此致謝。

參考文獻

大學入學考試中心（2019）。**大學入學考試素養導向命題簡介（稿）。**取自 https://www.ceec.edu.tw/files/file_pool/1/0J254620993850057887/1080529%E5%A4%A7%E5%AD%B8%E5%85%A5%E5%AD%B8%E8%80%83%E8%A9%A6%E7%B4%A0%E9%A4%8A%E5%B0%8E%E5%90%91%E5%91%BD%E9%A1%8C%E7%B0%A1%E4%BB%8B.pdf

[College Entrance Examination Center. (2019). *Introduction of literacy-oriented item development for college entrance examination (Draft)*. Retrieved from https://www.ceec.edu.tw/files/file_pool/1/0J254620993850057887/1080529%E5%A4%A7%E5%AD%B8%E5%85%A5%E5%AD%B8%E8%80%83%E8%A9%A6%E7%B4%A0%E9%A4%8A%E5%B0%8E%E5%90%91%E5%91%BD%E9%A1%8C%E7%B0%A1%E4%BB%8B.pdf]

大學入學考試中心（2020）。**111 大考素養導向命題方向與試題示例。**取自 https://www.ceec.edu.tw/files/file_pool/1/0K126642312488077919/20200504.pdf

[College Entrance Examination Center. (2020). *Guideline and item examples for literacy-oriented items in 2022 college entrance examination.* Retrieved from https://www.ceec.edu.tw/files/file_pool/1/0K126642312488077919/20200504.pdf]

王文中（1996）。幾個有關 Rasch 測量模式的爭議。**教育與心理研究，19，**1-26。

[Wang, W. (1996). Some controversial issues about the Rasch measurement model. *Journal of Education & Physchology, 19,* 1-26.]

王文中（1997）。測驗的建構：因素分析還是 Rasch 分析？**調查研究，3，**129-166。

[Wang, W. (1997). Constructs of a test: Factor analysis or Rasch analysis? *Survey Research, 3,* 129-166.]

余民寧（2009）。**試題反應理論及其應用。**台北：心理。

[Yu, M. N. (2009). *Item Response Theory and its application.* Taipei: Psychological Publishing.]

余民寧（2020）。**量表編制與發展：Rasch 測量模型的應用。**新北：心理。

[Yu, M. N. (2020). *Scale building and development: Application of Rasch measurement models*. New Taipei: Psychological Publishing.]

佘曉清、黃莉郁、蘇怡蓓（2017）。台灣學生科學素養的表現。載於佘曉清、林煥祥（主編），**PISA 2015 臺灣學生的表現**（頁 23-80）。台北：心理。

[She, H. -C., Huang, L. -Y., & Su, Y. -B. (2017). Taiwan student scientific literacy performance. In H. -C. She & H. -S. Lin (Eds.), *Taiwan student performance on PISA 2015* (pp. 23-80). Taipei: Psychological Publishing.]

林小慧、林世華、吳心楷（2018）。科學能力的建構反應評量之發展與信效度分析：以自然科光學為例。**教育科學研究期刊，63**（1），173-205。

[Lin, H. -H., Lin, S. -H., & Wu, H. -K. (2018). Developing and validating a constructed-response assessment of scientific abilities: A case of the optics unit. *Journal of Research in Education Sciences, 63*(1), 173-205.]

林春煌（2019）。素養導向教學與評量—以高中物理克卜勒行星定律為例。**中等教育，70**（3），114-122。

[Lin, C. -H. (2019). Literacy-oriented teaching and assessment: A case study of the high school physics Kepler' law. *Secondary Education, 70*(3), 114-122.]

林蓓伶、潘昌志、蘇少祖、陳柏熹（2018）。十二年國教國中階段自然科學領域素養導向評量試題之開發與初探。**教育科學研究期刊，63**（4），295-337。

[Lin, P. -L., Pan, C. -C., Su, S. -Z., & Chen, P. -H. (2018). Development of assessments for scientific literacy based on curriculum guidelines for 12-year basic education in science domains. *Journal of Research in Education Sciences, 63*(4), 295-337.]

洪碧霞、吳裕益、林哲彥、葉千綺（1992）。**大學入學考試題目分析時 IRT 模式選擇之初探。**台南：臺南師範學院測驗發展中心。

[Hong, B. -X., Wu, Y. -Y., Lin, Z. -Y., & Yeh, Q. -X. (1992). *Exploring IRT model selection in item analysis of college entrance examination.* Tainan: Test Development Center in National University of Tainan.]

郭伯臣、吳慧珉、陳俊華（2012）。試題反應理論在教育測驗上之應用。**新竹縣教育研究集刊，12，**5-40。

[Kuo, B. -C., Wu, H. -M., & Chen, C. -H. (2012). Application of IRT in educational measurement. *Journal of Educational Research Hsinchu County, 12,* 5-40.]

教育部（2018）。**十二年國民基本教育課程綱要國民中小學暨普通型高級中等學校—自然科學領域。**台北：作者。

[Ministry of Education. (2018). *Guidelines for the 12-year basic education curricula: Natural science.* Taipei: Author.]

陳彥君、洪素蘋（2019）。中學生現代公民核心素養之測驗編製。**教育實踐與研究，32**（2），39-79。

[Chen, Y. -C., & Hung, S. -P. (2019). Development of a modern citizen core literacy test for middle school students. *Journal of Educational Practice and Research, 32*(2), 39-79.]

陳柏熹（2011）。**心理與教育測驗：測驗編製理論與實務。**新北：精策教育。

[Chen, P. -H. (2011). *Psychological and educational measurement: Theory and practice of measurement building and development.* New Taipei: Besteam Education.]

陳泰然（2015）。**12 年國教自然科學領域課綱：科教老師自我實現的契機。**取自 https://slidesplayer.com/slide/11196843/

[Chen, T. -J. (2015). *Guidelines for the 12-year basic education curricula: Natural science: An opportunity for science education teachers' self-fulfillment.* Retrieved from https://slidesplayer.com/slide/11196843/]

張貴琳（2013）。青少年線上閱讀素養評量工具之發展。**教育實踐與研究，26**（2），29-66。

[Chang, K. -L. (2013). The development of adolescent online reading literacy assessments. *Journal of Educational Practice and Research, 26*(2), 29-66.]

盧政良、林蓓伶、何興中（2019）。2018 全國高中物理探究實作競賽初賽試題測驗分析研究。**物理教育學刊，20**（2），21-35。

[Lu, C. -L., Lin, P. -L., & Ho, H. -C. (2019). The item analysis study for the 2018 preliminary competition of inquiry and practice competition for high school in Taiwan. *Chinese Physics Education, 20*(2), 21-35.]

盧姮倩(2013 年 4 月 18 日)。類 PISA 喊卡家長轟浪費 8 個月　北市教局:浪費什麼？。**ETtoday 新聞雲。**取自 https://www.ettoday.net/news/20130418/194386.htm

[Lu, H. -G. (2013, April 18). *Stop PISA-type tests. Parents complained: Wasting 8 months. Department of education, Taipei city government said: What be wasted?* Retrievd from ETtoday Cloud Web site: https://www.ettoday.net/news/20130418/194386.htm]

Adeleke, A. A., & Joshua, E. O. (2015). Development and validation of scientific literacy achievement test to assess senior secondary school students' literacy acquisition in physics. *Journal of Education and Practice, 6*(7), 28-42.

Bybee, R. W. (1997). *Achieving scientific literacy: From purposes to practices*. Portsmouth, NH: Heinemann.

Bybee, R. W., McCrae, B., & Laurie, R. (2009). PISA 2006: An assessment of scientific literacy. *Journal of Research in Science Teaching, 46*(8), 865-883.

DeVellis, R. F. (2017). *Scale development: Theory and applications* (4th ed.). Thousand Oaks, CA: Sage.

Gonzalez, E., & Rutkowski, L. (2010). Principles of multiple matrix booklet designs and parameter recovery in large-scale assessments. *IEA-ETS Research Institute Monograph, 3,* 125-156.

Hurley, B. P., & Henry, M. P. (2015). Using a disciplinary literacy framework to teach high school physics: An action research study. *I.E.: Inquiry in Education, 7*(1), Article 3. Retrieved from: http://digitalcommons.nl.edu/ie/vol7/iss1/3

Kind, P. M. (2013). Conceptualizing the science curriculum: 40 years of developing assessment frameworks in three large-scale assessments. *Science Education, 97*(5), 671-694. https://doi.org/10.1002/sce.21070

Körhasan, N. D., & Gürel, D. K. (2019). Student teachers' physics knowledge and sources of knowledge to explain everyday phenomena. *Science Education International, 30*(4), 298-309.

Kuo, C. Y., Wu, H. -K., Jen, T. H., & Hsu, Y. -S. (2015). Development and validation of a multimedia-based assessment of scientific inquiry abilities. *International Journal of Science Education, 37*(14), 2326-2357.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

McCoubrie, P. (2004). Improving the fairness of multiple-choice questions: A literature review. *Medical Teacher, 26*(8), 709-712.

National Research Council [NRC]. (1996). *National science education standards.* Washington, DC: National Academies Press.

Nunnally, J. C. (1978). *Psychometric theory* (2nd ed.). New York, NY: McGraw-Hill.

Organisation for Economic Co-operation and Development [OECD]. (2017). *PISA 2015 science framework. In PISA 2015 assessment and analytical framework: Science, reading, mathematic, financial literacy and collaborative problem solving* (revised edition). Paris, France: Author.

Reckase, M. D. (1979). Unifactor latent trait models applied to multifactor tests: Results and implications. *Journal of Educational Statistics, 4*(3), 207-230.

Stelmack, J., Szlyk, J. P., Stelmack, T., Babcock-Parziale, J., Demers-Turco, P., Williams, R. T., & Massof, R. W. (2004). Use of Rasch person-item map in exploratory data analysis: A clinical perspective. *Journal of Rehabilitation Research & Development, 41*(2), 233-241.

Wilson, M. (2005). *Constructing measures: An item response modeling approach.* Mahwah, NJ: Lawrence Erlbaum Associates.

Wu, M., & Adams, R. (2007). *Applying the Rasch model to psycho-social measurement: A practical approach.* Melbourne, Australia: Educational Measurement Solutions.

投稿收件日： 2020 年 10 月 14 日
第 1 次修改日期： 2021 年 2 月 5 日
第 2 次修改日期： 2021 年 4 月 30 日
接受日： 2021 年 5 月 17 日