國立臺灣師範大學教育心理與輔導學系
教育心理學報，2019，50 卷，3 期，473-502 頁
DOI：10.6251/BEP.201903_50(3).0005

科學探究能力評量之標準設定與其效度檢核*

林小慧
國立台灣師範大學
教育學院

吳心楷
國立臺灣師範大學
科學教育研究所

本研究係以臺灣大台北地區 605 位 11 年級學生接受科學探究能力評量施測所蒐集的實徵資料，以達到兩項研究目的。其一為依據待加強、基礎、精熟三個等級之標準表現描述，設定科學探究能力評量之標準，其二則從內部、過程及外部等多元效度證據來源，檢核 Bookmark 法進行科學探究能力標準設定的適切性及有效性。研究結果顯示，本研究科學探究能力的標準設定可獲得過程效度證據的支持。其次，內部效度評估結果顯示，14 位標準設定成員在第一輪到第二輪之各表現等級的標準誤均在可接受範圍（SE < 0.12），表示成員內標準設定結果檢具可靠性。另以二輪決斷分數中位數之樣本平均數的標準誤評估標準設定方法內的一致性，結果顯示各表現等級的標準誤均在可接受範圍（SE < 0.12），表示標準設定方法內的結果相當一致。再者，以獨立樣本 *t* 檢定進行標準設定成員間一致性的考驗，分析結果顯示不同群組成員所設定的決斷分數均未達顯著差異。此外，標準設定極端值的監控結果發現，僅有少數極端值出現，故而對於整體決斷分數的影響甚微。因此，本研究科學探究能力標準設定可獲得內部效度證據的支持。最後，本研究以群聚分析標準設定，透過探討 Bookmark 法所得決斷分數之輻合效度，結果顯示二種標準設定法將學生分為三個表現等級之相關係數達顯著水準，表示在判斷表現等級有相當程度的一致性。另利用區別分析檢核標準設定的一致性，分析結果顯示，Bookmark 法在「觀察與定題」、「計畫與執行」、「分析與發現」及「推理與論證」整體分類一致性依序為 79.50%、86.00%、100.00%、89.90%，可見 Bookmark 標準設定法所得出的決斷分數在各表現等級分類之區別力相當高，可獲得外部效度證據的支持。綜合以上證據，研究結果顯示經由 Bookmark 法所設定之科學探究能力標準適切而且有效。

關鍵詞：科學探究能力、效度驗證、標準設定、Bookmark 標準設定法

* 本篇論文通訊作者：吳心楷，通訊方式：hkwu@ntnu.edu.tw。

臺灣國家教育研究院於 105 年公佈十二年國民基本教育自然科學領域課程綱要草案，此新版的自然科課綱預計於 108 年度實施（以下簡稱 108 課綱）。在新課綱中，學生在科學的學習成就，除了過去強調的「核心科學概念內容」之外，同時重視學生在「探究能力」和「科學的態度與本質」兩方面的「學習表現」，以因應未來其個人或社會發展的需要。針對學習重點的調整，教育部進一步規劃新課程，故而新高中課程分配架構，除延續既有物理、化學、生物及地科共同（必修）與進階（選修）課程外，另新規劃「自然科學探究與實作」課程共計四學分，以培養學生科學探究能力。由於 108 課綱的課程變革以及強調的學習表現，與長久以來自然科教師所熟悉的教學實務不盡相同，因此可以預見在未來新課綱正式實施時，高中自然科教師勢必面臨「探究與實作」的課程、教學與評量的挑戰。特別是探究實作的評量，其所需評測的能力，諸如發現問題、認識問題、問題解決、提出結論及表達溝通等，尤為顯得格外重要。

過去十年，國內學者已研發不少探究導向的學習模組（Hsu, Chang, Fang, & Wu, 2015; Hsu, Wu, & Hwang, 2008; Wu, 2010; Wu & Hsieh, 2006）、學習方案（陳慧娟，2015），以及科學評量（林小慧、林世華、吳心楷，2018），可協助科學教師推動探究與實作的教學。相比之下，探究能力評量的相關研究則較少。然而，教育評量的重要目的之一即在甄別學生的學習表現，若是缺乏適當的評量工具，教師可能無法了解學生的能力基點及需求，也無法評鑑其課程的有效程度。但是要如何比較才能區辨出好的表現？是將學生表現相比於團體內其他學生的表現，或是相比於預期的標準？此兩種不同的比較方式，即是目前常用的兩種參照方式：「常模參照」（norm referenced）和「標準參照」（standard referenced）。前者係植基於相對標準的模式，適於安置性或總結性評量（吳清山，2014），主要將所蒐集之學生評測資料的答題反應等化到同一量尺上，透過受試學生在評量表現的相對位置，決定決斷分數，並藉由百分比將學生分級。然而，在此種參照方式下，學生的表現等級會因為比較群體的不同，而可能被歸類在不同的水準，亦即學生的表現處於變動狀態，端視所比較群組的程度而定。基此，常模參照所選定的決斷分數若要具有意義，則前提必須滿足每次的受試母群相似，以及試題所植基的雙向細目表要相同的假設（Cizek & Bunch, 2007; Green, Trimble, & Lewis, 2003）。而標準參照較適用於診斷性評量和發展補救教學（吳清山，2014），且不因受試者表現而改變。因此，本研究研究目的之一為設定科學探究能力評量之標準：採用兼具選擇題型與建構題型之多媒體電腦化評量（Multimedia-based Assessment of Scientific Inquiry Abilities, MASIA），設定科學探究能力評量之標準。

目前國內關於標準設定的研究，多半聚焦在各學科之學習成就表現標準設定的效度檢核，諸如國語文（曾建銘、王暄博，2012a）、英文科（謝名娟、謝進昌、林世華，2013）、數學科（吳宜芳、鄒慧英、林娟如，2010），以及社會科（曾建銘、王暄博，2012b）。關於科學探究能力這類能力取向之標準設定，尚未有相關研究的探討。再者，本研究採用多媒體電腦化評量進行科學探究能力的檢測，有別於過去研究所使用的紙本測驗，在進行標準設定時，除了提供標準設定成員有關評量的紙本資料以外，亦在現場備有電腦以利成員能夠實際觀測學生接受評量時所看到的試題介面。本研究所提供的標準設定流程及結果，可供未來在能力導向或電腦化評量相關研究做為參考。

其次，目前常用的標準設定方法為 Angoff 法和 Bookmark 法（又稱書籤法）。考量適用於多題型評量的優勢，本研究將採 Bookmark 法設定科學探究能力表現標準。然而，儘管 Bookmark 法便利使用的特性越來越受到證明及肯定（Hambleton, Jaeger, Plake, & Mills, 2000），但仍缺乏關於信度與效度的研究文獻，致難有充足論據予以支持（Peterson, Schulz, & Engelhard, 2011）。另有研究指出（Green et al., 2003; Reckase, 2006），Bookmark 法所得出的決斷分數會有低於其他設定方法的現象，如 Yin 與 Schulz（2005）發現 Bookmark 法低於 Angoff 方法，造成負向偏誤（過低）。有鑑於此，本研究第二項研究目的，即依據 Kane（1994）所提出標準設定效度評估的觀點，從內部、過程及外部等多元效度證據來源，來檢核以 Bookmark 法進行科學探究能力標準設定的適切性及有效性。其中，外部效度部分，本研究另採統計程序之群聚分析（cluster analysis）進行標準設定，透過探討 Bookmark 法所得決斷分數之輻合效度（convergent validity）的證據，瞭解所得出決斷分數的適切程度。

一、標準設定的內涵

Sturmberg 與 Hinchy（2010）指出，通過或不通過是一個複雜的決定，必須要有一個最低能力水準的定義。Cizek（2006）指出標準設定（standard setting）係指事先設定好二個或多個表現等級的評定準則，並依據準則建立一個或多個決斷分數的歷程。換句話說，標準設定係指為已發展之評量建立一系列判斷標準的過程，用以定義學習成就水準或專業程度的方法。Cizek、Bunch（2007）強調決斷分數係為評估學生是否達到既定成就或專業水準的依據，可將受試者區分二個或多個類別，例如通過、不通過，或者基礎、精熟及進階。

Reckase（2000）認為標準設定應將應實務與方法列入考量，並融入決策者與研究者的理念需求或元素。Loomis、Bourque（2001）指出標準設定方法的原則，大多強調多元、融貫與適切性，例如美國全國教育進步測驗（National Assessment of Educational Progress，以下簡稱 NAEP）判斷標準設定方法適切性的六大準則，包括：與 NAEP 計分、量尺與分析技術一致；擁有完善暨可驗證之決斷分數計算的統計歷程；能將標準設定成員之主觀判斷反應在 NAEP 量尺上；能充分發揮可獲得的資料；能將政策的決定權留給執政者；檢具清晰、簡要、易於解釋與實用性。

二、標準設定的方法

目前廣被運用之標準設定方法，大致可區分為受試者中心模式（examinee-centered model）與測驗中心模式（test-centered model）（Kane, 1994）。受試者中心模式則由學科專家依據受試者的答題表現，決定通過分數，諸如對照組法（Livingston & Zieky, 1989），即為隨機取樣被判定「通過」及「不通過」二組，再畫出二組分數與人數的分布圖，進而從二組分數重疊區選擇通過分數。但此模式訂定之標準會因受試者表現而改變，因此本研究採測驗中心模式，係由學科專家針對試題特性判斷最低能力受試者的答題表現，並求得該群受試者可能的得分，以作為通過分數，諸如 Angoff 法（Angoff, 1984）、Ebel 法（Ebel & Frisbie, 1986）、Nedelsky 法（Nedelsky, 1954）、Jaeger 法（Jaeger, 1982），以及 Bookmark 法（Lewis, Mitzel, & Green, 1996）等。

（一）Angoff 法

Angoff 法係為目前學界廣為運用的標準設定方法之一，Angoff（1971）要求參與標準設定成員判斷最低能力表現者於每道試題的答對率，並求出所有標準設定成員的平均值，作為精熟標準門檻。有鑑於原始 Angoff 法面對測驗試題眾多或需判斷受試者水準較多時，容易產生評定歧異暨不易達成評定共識，是以衍生改良式選擇型 Angoff 法及 Yes/NoAngoff 法。前者係將 7 種判定的答對率（5%、20%、40%、60%、75%、90%、95%）提供給標準設定成員，請其選擇最低能力受試者答對的百分比，並加總每位成員各題的答對率，最後求出所有成員之答對百分比總和的平均值，作為精熟標準門檻（Berk, 1986）。後者則是要求成員逐題判斷最低能力受試者於每道試題答對與否，若能答對寫「Yes」，不能答對則寫「No」，接著計算每位成員判定「Yes」於整份測驗所佔比例的平均值，以作為精熟標準門檻，此法有利於減少標準設定成員間的評定變異（Impara & Plake, 1997）。

（二）Bookmark 法

NAEP 長久以來均使用 Angoff 法來建立決斷分數，直到 2005 年國家評量管理委員會（National Assessment Governing Board）開始評估 Bookmark 法之信度與效度，從那時候起 NAEP 便開始採用 Bookmark 法並逐漸取代 Angoff 法，理由是相較於 Angoff 法，小組成員判斷決斷分數會更為可靠，並有標準設定時間較短暨成本較低的優勢（Peterson, et al., 2011）。Perie 指出截至 2005 年，美國已有 31 州使用 Bookmark 法進行標準設定，成為使用頻率最多的標準設定方法（引自 Karantonis & Sireci, 2006）。

Bookmark 法的實施程序，首先準備一份經由試題反應理論（item response theory，簡稱 IRT）所估計之試題難度，並且由易至難排序好的試題卷（ordered item booklet，以下簡稱 OIB），依照一

頁一試題，並包含題目內容、選項、計分規準（scoring rubrics）等訊息。其次，召集與培訓標準設定成員，並使其熟悉內容標準與表現等級的描述。Huynh（2006）指出，二元計分題目的訊息量在 $p = .67$ 達到最大，亦即當學生答對機率為 .67 時，能力估計誤差最小。因此，本研究要求標準設定成員依據 Mitzel、Lewis、Patz 和 Green（2001）所建議 67%的反應機率，判斷最低能力受試者可達 67%答對率的題目，並將書籤放置在該試題位置，作為不同表現水準的切截點。最後，依據每位標準設定成員將各表現等級書籤所放置試題的難度，在 67%答對率的條件下，進行能力參數（examinee's ability, θ）的轉換，並求出平均能力值，再轉換成原始分數，此即為決斷分數。Cizek（2006）指出，Bookmark 法的優點在於作法較易了解與容易執行，可避免 Angoff 法逐題檢視及評定之耗時費力的疑慮，亦有適用於建構題（constructed-response items）與選擇題（selected-response items）兼具之評量的優勢。

綜上所述，本研究採用 Bookmark 法，主要考量此法在操作上較易了解與容易執行的優勢，不僅可避免 Angoff 法逐題檢視及評定之耗時費力的疑慮，並可適用於建構題與選擇題兼具的評量卷。本研究蒐集臺灣大台北地區 11 年級學生接受科學探究能力評量施測資料，結合 IRT 分析技術估計試題難度所得之訊息，提供給標準設定成員，判斷最低能力受試者 67%答對率試題暨估算所對應的能力值，求得平均能力值後，進而轉換成原始分數，即得科學探究能力評量報表各表現等級的決斷分數。

然而，儘管 Bookmark 法便利使用的特性越來越受到證明及肯定（Hambleton, et al., 2000），但仍缺乏關於信度與效度的研究文獻，致難有充足論據予以支持（Peterson, et al., 2011）。另有研究指出 Bookmark 法所得出的決斷分數會有低於其他設定方法的現象（Green et al., 2003; Reckase, 2006）。因此，本研究將透過多項效度證據，來瞭解 Bookmark 法所得出之決斷分數的適切程度。

三、標準設定的程序

為回應第一項研究目的，除了採用 Bookmark 法之外，本研究應用 Cizek、Bunch（2007）所建議執行標準設定的程序，包括釐清測驗目的、選擇標準設定方法、訂定表現等級名稱並建置表現等級的描述、選擇及訓練標準設定成員、選擇標準設定方法、提供回饋給成員、監控標準設定歷程等步驟，茲闡述如下：

（一）釐清測驗目的（identify/clarify purpose of the assessment）

測驗目的（如：形成性、診斷性、或總結性評量），不僅會影響測驗的型態、結構或特性，以及標準設定的目的與決斷分數的建置，亦會影響表現等級個數的決定暨表現標籤的命名。

（二）選擇標準設定方法（choose a standard-setting method）

Cizek、Bunch（2007）針對標準設定的選擇，提出六個重要因素，包括：（1）服膺測驗的目的；（2）呼應測驗所評測之知識、技能及能力的複雜水準；（3）符合測驗的格式，例如 Nedelsky 法適合選擇題型，Angoff 法則適合選擇或建構反應題型；（4）考量表現等級的個數，亦即為決斷分數；（5）考量可用資源的程度；（6）最後則是提出使用多元標準設定方法的可行性，若在資源有限情況下，徹底執行單一方法較執行二種或多種方法，卻不貫徹要來得好。

（三）訂定表現等級個數暨標籤命名（create performance level labels）

Cizek（2006）指出表現等級標籤（performance level labels）係用以辨識表現類別，如前所述，本計畫植基 NAEP 所提出之表現標準標籤命名藍圖，將本計畫科學探究能力評量之表現等級分為待加強、基礎及精熟三個等級。

（四）建置表現等級的描述（prepare performance level descriptions）

表現等級的描述（performance level descriptions，簡稱 PLDs）係針對特定等級關於表現的完整說明。Cizek、Bunch（2007）指出，完成表現標準個數的訂定暨標籤的命名後，研究者必須接續為各表現等級之實質內涵進行闡述，進而具體界定基礎、精熟及進階等表現等級的描述。

（五）確認關鍵概念（form key conceptualizations）

所有標準設定方法都有必要為標準設定成員，形成進行評判所需要的概念，此即為成員在標準設定歷程中重新審視的關鍵參考，並且有助於解釋產生決斷分數的意義，這些關鍵概念包括：答對機率、表現等級、PLDs、邊界受試者（borderline examinee）。Cizek、Bunch（2007）進一步以 Angoff 法為例，標準設定成員必須逐題審評，形成對關鍵概念的共識，並估計最低能力受試者正確回答該題的機率。Giraud、Impara 與 Plake（2005）顯示有關目標受試者的特性定義或者是小組成員的討論，都會影響評審者的判斷，因此標準設定成員概念化假想受試者的能力，即為標準設定程序成功的關鍵。

（六）選擇及訓練標準設定成員（select and train standard-setting participants）

為使參與成員瞭解與熟悉標準設定的目的和技術，選擇適切之標準設定成員，並提供培訓課程，是提升標準設定效度的重要環節。同時標準設定成員也應具有母群代表性，包括領域代表性暨各區域的代表性，例如研究人員、評量設計人員及學科教師等。

（七）提供回饋給標準設定成員（provide feedback to participants）

為有效協調成員凝聚共識或具體化特定概念，在各標準設定階段，應該提供參照資料回饋給標準設定成員，包括事實訊息、影響訊息，以及常模參照訊息（Cizek & Bunch, 2007）。其中，事實訊息（reality information）主要在協助成員覺知自身判斷的準確性，如試題難度、鑑別度等。另外，影響訊息（impact information）則提供成員設定決斷分數所可能造成的影響訊息，諸如各水準通過人數的百分比，成員可就此訊息進行決斷分數對於社會大眾的觀感與接受程度的討論。而常模訊息（normative information）則是用以協助成員覺知自身與其他成員判斷的歧異，包括每位成員所設定之決斷分數及其極端值的分布、平均數、中位數、標準差等。以 NAEP 為例，其所提供之回饋訊息有：題本、試題難度、成員判定位置及一致性訊息等（Reckase, 2001），而提供這些回饋訊息的目的，係使標準設定成員判定結果的品質能夠達到最佳化，並提升成員間判定的一致性（Loomis, 2000）。

（八）監控標準設定歷程（evaluate the standard-setting process）

Cizek（2006）認為標準設定的監控，包括方法的選擇、成員的招募和訓練，以及貫徹執行程序的規劃。其中，方法選擇的決策，應該綜合考量評量目的、格式、可使用資源，以及欲評核知識技能的水準和個數；標準設定成員則須有母群代表性，同時應接受培訓致使了解和熟悉標準設定的技術與目的；執行標準設定程序則應秉持嚴謹的態度，貫徹每個流程步驟，藉以確保標準設定的品質。

四、標準設定的效度驗證

由於本研究採用的 Bookmark 法，仰賴人為主觀的判斷，故而不可避免主觀判斷所產生的歧異，有必要透過實徵檢驗以建立表現標準的效度，即決斷分數的解釋效力。因此，效度驗證為判定標準設定的結果，是否檢具表現標準分類（待加強／基礎／精熟）的準確性、合理性與實務應用性，即為標準設定研究歷程的重要項目。決斷分數和 PLDs 的效度取決於過程各階段的執行品質和證據的評估。為達成本研究第二項研究目的，即檢核以 Bookmark 法進行科學探究能力標準設定的適切性及有效性，本研究採 Kane（1994）提出三種效度證據，包括過程證據（procedural evidence）、內部證據（internal evidence），以及外部證據（external evidence）。

（一）過程證據

過程證據著重標準設定過程的適當性暨執行程序的品質，是評估表現標準重要項目之一。本研究藉由檢核標準設定方法的選擇與執行、標準設定成員的選擇與訓練、標準設定成員的訊息回饋，以及成員判斷結果之聚斂程度作為過程效度的證據。

（二）內部證據

Raymond、Reid（2001）指稱有鑑於每位標準設定成員之專業、經驗和興趣的差異，故而成員間的判斷存在變異是可以預期跟理解的，因此內部效度證據係強調標準設定成員判斷結果是否檢

具穩定性與一致性。本研究藉由檢核標準設定成員內與成員間的一致性，以及標準設定方法內的一致性作為內部效度的證據。

（三）外部證據

效度的外部證據係強調標準設定方法間的一致性，或標準設定結果與其他相關規準的關係，包括學生相關學科的表現成績、學生答題表現的群聚分析，以及決斷分數之可行性與真實性的程度。本研究即採群聚分析法之結果做為外部證據。

實徵研究結果發現，以群聚分析所獲得的決斷分數，和經由專家討論所獲得的結果相當接近（Violato, Marini, & Lee, 2003），足見統計模式與人為判斷這兩種取向的標準設定方法間存在相當程度的一致性。群聚分析法係利用「距離」概念，進行變項間之相似性的分析，主要包括階層性群聚法（hierarchical clustering method）以及非階層性群聚法（nonhierarchical clustering method）兩類，前者係將資料層層反覆地進行分裂（由眾多群體逐次分裂成少數群體）或聚合（將少數群體逐次合併成成眾多群體）的歷程，後者則指在各階段分群過程中，將原有的集群打散，並重新形成新的集群，以 k 組平均法（k-means）為代表（Timm, 2002）。由於本研究將學生分為待加強、基礎及精熟三個能力等級，故本研究採用 k 組平均法（$k = 3$），以歐幾里德距離（Euclidean distance）代表個體間的距離，作為分組的依據。

方法

一、研究架構

標準設定流程如圖 1 所示，本研究首先發展符合 108 課綱探究能力之評量架構，依架構設計評量試題，並依序完成試題施測及試題難度估計。其次，召集六位領域專家學者及八位高中自然科領域教師，成立標準設定小組並進行評量標準的制訂，包括表現標準個數暨命名類別、各表現等級的陳述。再者，依據評量標準訂定評分規準暨確立各表現等級的「樣卷」及「樣卷說明」示例，藉以提供評分者評核範例的參考。接著依序召開二輪標準設定會議，進行標準設定。其中第二輪標準設定會議將提供成員第一輪設定結果的回饋訊息，並於會中共同討論之後，再進行第二輪標準設定。最後分析階段，包括將小組成員的設定結果進行分析暨轉換成決斷分數，以及進行標準設定的效度評估。

二、發展科學探究能力的標準卷

（一）評量範疇

本研究首先綜合課綱及文獻整理暨徵詢學科專家與現職教師的意見，發展符合 108 課綱探究能力之評量架構，並依此架構從已發展之評量系統挑選試題組成標準卷，共涵蓋「觀察與定題」、「計畫與執行」、「分析與發現」及「推理與論證」四個次能力，內容包括浮力、小鐵球、紙飛機、滾罐子、氣體體積、物體下落、下沈快慢等七個單元，共計 39 題（參照表 1）。

表 1　探究能力評量架構

探究能力次能力	子能力	內容	題數
（Q）觀察與定題 透過觀察周遭的事物和現象，察覺或訂定可被驗證的問題，並預測可能的答案。	提出預測	能透過先前的經驗、概念或觀察結果，來預測研究問題可能的答案。	8
	確認問題	辨識或提出與情境相符且可被驗證的研究問題。	3
（E）計劃與執行 依據問題辨認自變項與應變項、並選擇適當的工具或儀器來擬定實驗流程。	辨識變因	能辨認相關的自變項與應變項來擬定實驗流程。	8
	規劃實驗	能描述與制定實驗流程，並可驗證變項關係。	4
（A）分析與發現 分析資料數據找出變項之間的關係或趨勢，提出符應該關係或趨勢的科學主張。	分析資料	能挑選資料數據驗證變項關係，或將數據以另一種表徵方式呈現，以驗證變項關係。	6
	提出主張	能透歸納、演繹的方式辨識出資料的分佈趨勢，來形成可驗證的陳述或論點。	3
（R）推理與論證 運用適當的資料數據支持主張，並透過推論的過程來提出結論或解釋。	運用證據	能透過歸納找出正確數據，以支持主張。	5
	產生推理	將證據連結到主張，包含使用科學原則、概念或先前經驗進行推理，詮釋或推論資料的意義。	2
		合計	39

（二）評量格式與評量原則

本評量包括選擇題與建構題，前者又可分為二元計分及多元計分之次序型選擇題（ordered multiple choice，以下簡稱 OMC），亦即 OMC 的每個選項代表不同的學習層級，能夠反應學生能力屬於那個程度；後者則包括簡答題、填充題、勾選題、繪圖題及申論題，採用多元計分模式，依據學生的作答反應進行評分，評分原則包含「無法正確地……」評 0 分、「能部分正確地……」評 1 分、「能正確地……」評 2 分，未作答，則記為“.”。

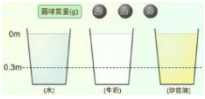
（三）評分規準的發展

本研究針對建構題與 OMC 題型發展評分規準，用以闡釋內容標準、表現標準，以及所評估能力表現的指引，茲以例題簡述評分規準如表 2。

（四）施測與難度估計

本研究首先從包含臺北市、新北市及基隆市 105 所公私立高中，依據 102 學年度入學考試的百分等級（percentile rank）區分為八層，並從各分層中隨機抽取一間學校作為抽樣的受試學校，再從各抽樣學校 11 年級中隨機抽取兩個班級進行施測，受試學生合計 605 位。本評量採團體施測，每位受試者均在電腦介面接受測驗，首先請其填寫基本資料，其次由研究者進行作答說明，最後正式進行施測，共計 50 分鐘。待施測資料蒐集後，先請評分者依據評分規準進行評分，依序進行試題內部一致性的檢驗、驗證性因素分析（confirmatory factor analysis, CFA），以及試題難度的估計。

表 2　評量例題與評分規準之示例

	探究能力－觀察與定題	次能力－提出預測
試題	【下沉快慢 1】 小華認為物體在液體中下沉快慢可能與液體密度有關。 針對這個看法，他準備體積相同但質量不同的圓球：20 克（g）、25 克（g）、30 克（g）。將這些圓球分別放入水、牛奶、沙拉油等液體中。測量圓球從液面下沉至深度 0.3 公尺（m）所需時間。 實驗裝置如下圖：  子題 1-1. 請你依據小華的想法，預測**液體密度與物體下沉速度**的關係，並選出正確答案。	
	我的預測（下拉式選單）：	
	○A. 液體密度越大，下沉時間越短 ○B. 液體密度越小，下沉時間越短 ○C. 液體密度與下沉時間無關	
評分標準	答案　C. 液體密度與下沉時間無關	
	答對　2 分	
試題	子題 1-2. 請你依據小華所預測**液體密度與物體下沉速度**的關係，並利用**科學知識**說明理由。	
	我的預測（下拉式選單）：	寫出影響或不影響的理由（填答）：
	○A. 液體密度越大，下沉時間越短 ○B. 液體密度越小，下沉時間越短 ○C. 液體密度與下沉時間無關	
評分標準	答案　1. 跟黏稠度比較有關係。 2. 密度不會影響下沉時間，只會決定浮沈。 3. 下沉時間和物體密度與液體密度差有關因為浮力造成的阻力大小不一樣。	
	答對　2 分	
評分標準	答案　1. 因為密度愈大物體愈難往下沉**[有提及密度對下沉難易的影響]**。 2. 液體密度比球大，則圓球越不容易下沉**[有比較液體密度與物體密度]**。 3. 密度越大會造成阻力使物體下沉變慢**[有以密度與阻力去解釋下沉快慢]**。 4. 因為只要液體密度較物體密度小物體較不會受阻**[有比較液體密度與物體密度]**。 5. 因為密度越大代表越濃稠所以下沉速度慢密度小下沉速度快**[以密度與濃稠度去解釋下沉快慢]**。 6. 越濃的沉的越慢[以密度與濃稠度去解釋]。 7. 液體密度小空間大（空隙多），阻力小。	
	答對　1 分	

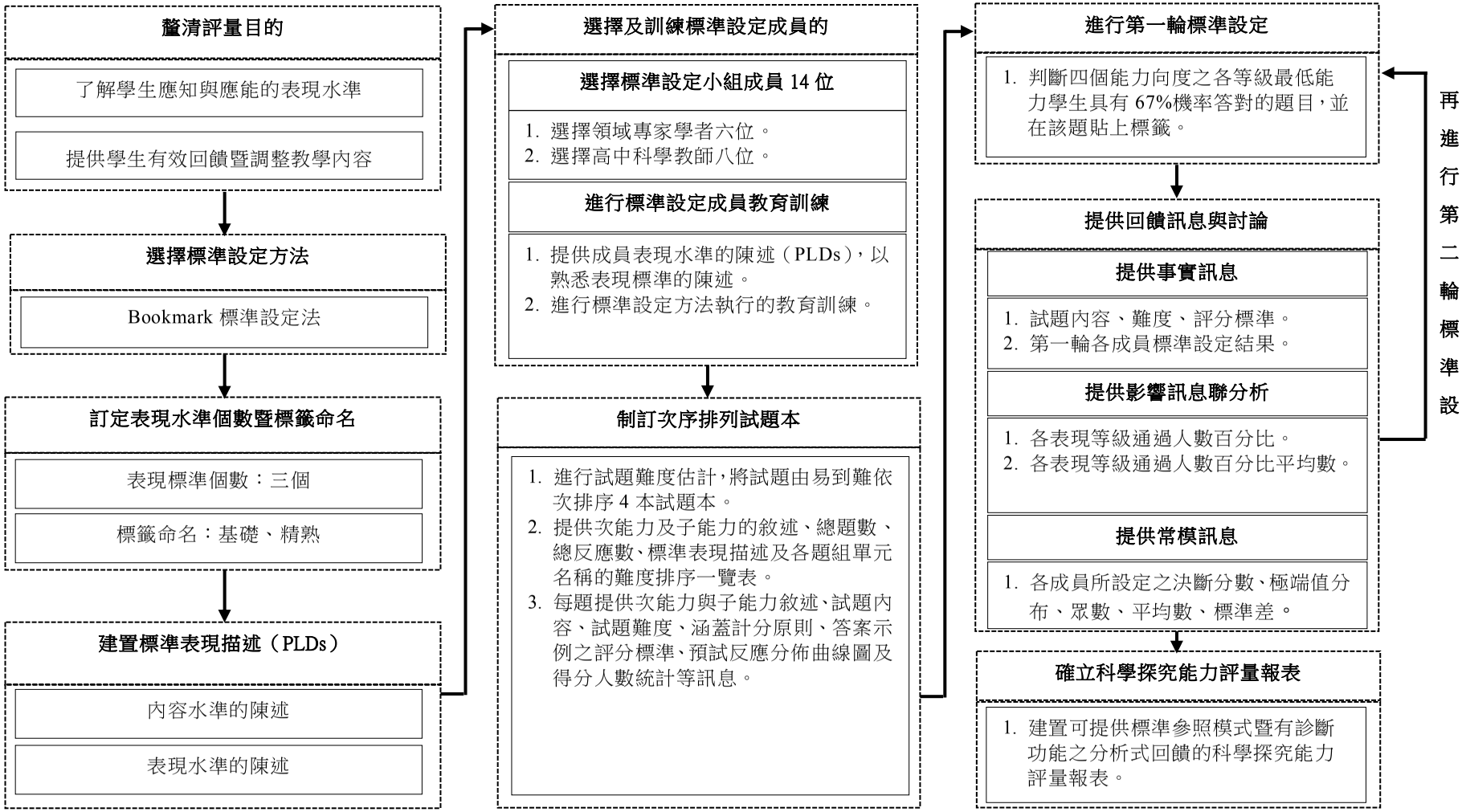

釐清評量目的
了解學生應知與應能的表現水準
提供學生有效回饋暨調整教學內容
選擇標準設定方法
Bookmark 標準設定法
訂定表現水準個數暨標籤命名
表現標準個數：三個
標籤命名：基礎、精熟
建置標準表現描述（PLDs）
內容水準的陳述
表現水準的陳述
選擇及訓練標準設定成員的
選擇標準設定小組成員 14 位
1. 選擇領域專家學者六位。
2. 選擇高中科學教師八位。
進行標準設定成員教育訓練
1. 提供成員表現水準的陳述（PLDs），以熟悉表現標準的陳述。
2. 進行標準設定方法執行的教育訓練。
制訂次序排列試題本
1. 進行試題難度估計，將試題由易到難依次排序 4 本試題本。
2. 提供次能力及子能力的敘述、總題數、總反應數、標準表現描述及各題組單元名稱的難度排序一覽表。
3. 每題提供次能力與子能力敘述、試題內容、試題難度、涵蓋計分原則、答案示例之評分標準、預試反應分佈曲線圖及得分人數統計等訊息。
進行第一輪標準設定
1. 判斷四個能力向度之各等級最低能力學生具有 67%機率答對的題目，並在該題貼上標籤。
提供回饋訊息與討論
提供事實訊息
1. 試題內容、難度、評分標準。
2. 第一輪各成員標準設定結果。
提供影響訊息聯分析
1. 各表現等級通過人數百分比。
2. 各表現等級通過人數百分比平均數。
提供常模訊息
1. 各成員所設定之決斷分數、極端值分布、眾數、平均數、標準差。
確立科學探究能力評量報表
1. 建置可提供標準參照模式暨有診斷功能之分析式回饋的科學探究能力評量報表。
再進行第二輪標準設

圖 1　標準設定流程

三、評量標準的制訂

本研究採用 Cizek、Bunch（2007）所提出標準設定程序架構，召集學科專家制訂評量標準（參照表 3）。

表 3　科學探究能力之標準表現描述

探究能力次能力	表現等級（表現標準）	標準表現描述（內容標準）
觀察與定題（Q）	待加強	1. 提出的預測與研究問題無關。 2. 能辨識或提出可被驗證的研究問題，但該問題與所提供的訊息不符。
	基　礎	1. 能利用先前經驗、概念或觀察結果，對研究問題可能的答案提出預測，但預測產生的理由無法說明或說明不清楚。 2. 能辨識或提出與所提供的訊息相符且可被驗證的研究問題，但僅是描述現象的問題。
	精　熟	1. 能利用先前經驗、概念或觀察結果，對研究問題可能的答案提出預測，並說明預測產生的理由。 2. 能辨識或提出與所提供的訊息相符且可被驗證的研究問題，並明確指向某種變項間的關係，即具因果性、關係性問題。
計劃與執行（E）	待加強	1. 無法辨認自變項與應變項。 2. 設計的流程可行，但是涉及的變項錯誤或無關。
	基　礎	1. 使用的變項不完整，只有操縱自變項，或只有控制其他條件。 2. 設計流程可行，但所獲取的資料不足以驗證所涉及的變項關係。
	精　熟	1. 能在控制其他條件的狀態下，操縱單一自變項來進行實驗。 2. 設計的流程可行，且所獲取資料足以驗證所涉及的變項關係。
分析與發現（A）	待加強	1. 能挑選資料數據來驗證變項關係，但與研究問題無關，或能將資料數據以表格方式來呈現變項間的關係。 2. 做出與研究問題無關的主張，該主張未涵蓋資料趨勢或變項關係。
	基　礎	1. 挑選的資料數據只能驗證研究問題部分的變項間關係，或能將資料數據以圖形的方式來呈現變項間的關係。 2. 做出一個不完整的主張。該主張只涵蓋部分資料趨勢或部份變項關係。
	精　熟	1. 挑選完整的資料數據能驗證研究問題的變項間關係，或能以模型、數學算式、方程式來呈現變項間的關係。 2. 做出一個完整的主張。該主張可涵蓋資料趨勢或變項關係。
推理與論證（R）	待加強	1. 提供不適當的資料數據做為證據（即證據無法支持主張）。 2. 使用錯誤的方式來推論證據與主張的關係。
	基　礎	1. 提供做為證據的資料數據僅能部分支持主張。 2. 提供連結主張和證據的推論，但連結的關係不完整或部分錯誤。
	精　熟	1. 提供做為證據的資料數據能支持主張。 2. 提供連結證據到主張的推論，並明確指出證據與主張的關係。

（一）決定標準設定方法、訂定表現等級個數及建置標準表現描述（PLDs）

本研究發展「科學探究能力評量」，旨在提供高中科學教師評估學生「自然科學探究與實作」課程之學習狀況的參考，為避免師生對評量結果詮釋和理解的困難，因此採用標準參照模式，用以評定及描述學生是否達到預期的目標或程度。有鑑於科學探究能力包括「觀察與定題」、「計畫與執行」、「分析與發現」、「推理與論證」四個能力向度，為避免標準設定成員必須就四個次能力進行標準設定的任務負擔，暨考量前導研究已有大量施測資料(Wu, Kuo, Jen, & Hsu, 2015)，可茲結合 IRT 技術估計試題難度，故而採用執行容易暨適用於建構與選擇題型兼具之評量卷的 Bookmark 標準設定法，請成員判斷最低能力受試者 67%答對率試題，

之後估算所對應的能力值，暨求得平均能力值後，進而轉換成原始分數，即得科學探究能力評量各表現等級的決斷分數。

本研究首先進行科學探究能力表現的分析，了解高中生在修習「自然科學探究與實作」課程後，所應具備的能力。其次，將表現等級分為待加強、基礎及精熟三個等級，並依據評量標準架構，進行實質內涵的闡述，具體界定待加強、基礎及精熟表現的描述。其中評量標準架構包括內容標準及表現標準二部分，茲說明如下：

1. 內容標準

內容標準意指界定學生在學習歷程中應學會的知識與技能（Hambleton, 2001）。以本研究制訂科學探究能力內容標準為例，即是將 108 課綱中的四項探究能力作為主要檢核指標，包括「觀察與定題」、「計畫與執行」、「分析與發現」及「推理與論證」。每個指標之下則各發展二個子能力，每個子能力的重點意涵，係植基 108 課綱探究能力之「學習表現」的描述。

2. 表現標準

表現標準係用以說明學生在經歷一個學習階段後，所應學會知識與技能的程度（Linn & Herman, 1997），包括表現等級與表現等級描述（PLDs）。由於學生表現會有程度上的差異，故而必須藉由表現等級描述來加以區分，因此，本研究為提供教師更多學生表現的訊息，並可能未來將標準卷同時應用於國高中，因此將科學探究能力的學習表現區分為精熟、基礎及待加強三個等級，同時針對各個表現等級建置 PLDs，以界定不同等級應具備的表現。

（二）制訂次序排列試題本（OIB）

本研究首先以 IRT 技術估計試題難度，進行部分給分模式（partial credit model）之試題難度的估計，並依照試題難度由易到難依次排序「觀察與定題」、「計畫與執行」、「分析與發現」、「推理與論證」四本 OIB。每本試題卷的第一頁，均提供該次能力及子能力的敘述、總題數、總反應數、標準表現描述，以及各題組單元名稱的難度排序一覽表。此外，OIB 係為一頁一試題的設計，每頁包括試題內容、試題難度、涵蓋計分原則、答案示例之評分標準，以及預試反應分佈曲線圖和得分人數統計等訊息。

（三）選擇及訓練標準設定成員

1. 召集標準設定小組

本研究標準設定小組成員（表 4），包括八位（57%）高中自然科領域教師、六位（43%）領域專家學者，共計十四位（參照表 4）。其中，高中自然科領域教師包括自然科領域教師五位（36%）、自然科領域教科書編（審）教師 1 位（7%）、自然科領域學科中心教師一位（7%）及具備自然科領域教學經驗的行政人員一位（7%）。另領域專家學者則指熟悉高中自然科領域之課程與教學五位（36%），以及評量架構的大學教授或研究員一位（7%）。另在性別分布，男性四位（29%），女性十位（71%）。

表 4　標準設定成員專業領域背景

領域類別	職務領域	性別（%）				人數（%）	
		男		女			
自然科領域教師	教學教師	2	（14.00）	5	（36.00）	7	（ 50.00）
	行政人員	0	（ 0.00）	1	（ 7.00）	1	（ 7.00）
領域專家學者	大學教授	1	（ 7.00）	4	（29.00）	5	（ 36.00）
	研 究 員	1	（ 7.00）	0	（ 0.00）	1	（ 7.00）
	合　　計	4	（29.00）	10	（71.00）	14	（100.00）

2. 培訓標準設定成員

本研究成立標準設定小組以後，首先寄發關於標準設定會議的相關資料，包括本研究標準設定的目的與進程、標準設定方法的理論與執行說明、各表現等級之標準表現描述、標準卷的評量架構，以及標準設定的會議流程，裨益成員了解所要執行的任務。其次於會議當中，研究者再次說明會議的目的、流程，以及簡要介紹標準設定方法的材料與執行步驟，並請成

員逐一檢視試題卷暨提出討論，最後再請成員進行 Bookmark 法之標準設定，判斷基礎及精熟之最低能力學生具 67%答對率的題目，並將書籤放置在試題位置。

（四）提供標準設定成員回饋訊息

為有效協調成員凝聚共識，本研究在第一輪標準設定階段，提供各成員試題難度由易到難排序的試題卷（OIB），內容包括次能力及子能力的敘述、標準表現描述、試題內容、試題難度、評分標準，以及預試反應分佈曲線圖及得分人數統計等訊息。另在第二輪標準設定階段，除了提供第一輪標準設定結果之事實訊息外，也提供各表現等級通過人數百分比及其平均數之影響訊息，以及各成員所設定之決斷分數與極端值的分布、眾數、平均數、標準差等常模訊息，以提升成員標準設定結果的品質和一致性。

（五）標準設定結果轉換決斷分數

本研究將各成員針對「觀察與定題」、「計畫與執行」、「分析與發現」、「推理與論證」等次能力，判斷最低能力受試者 67%答對率之試題並加以標籤設定的結果，估算所對應的能力值。有鑑於 IRT 係在 50%的答對機率（ $P=0.5=\frac{e^{(\theta-\beta)}}{1+e^{(\theta-\beta)}}$ ）進行能力估計，因此必須再將 50%答對率條件下所估出的能力值（θ），轉換為 67%答對率時的能力值（ θ'，參照公式 1.1）。最後求得平均能力值後，將達到平均能力值學生的作答反應，估算答對反應數，進而求得平均數，再轉換成原始分數，即為科學探究能力評量報表各表現等級的決斷分數。

$$P=0.67=\frac{e^{(\theta'-\beta)}}{1+e^{(\theta'-\beta)}}\rightarrow\theta'=(\beta)+\ln\left[\frac{0.67}{0.33}\right]=(\theta)+\ln\left[\frac{0.67}{0.33}\right] \qquad 1.1$$

四、標準設定的效度評估

本研究依據 Kane（1994）所提出效度評估向度的建議，依序就過程、內部及外部三種效度證據進行效度檢核（參照表 5）。

表 5　標準設定效度評估一覽表

效度證據	效度評估來源	評估說明
過程效度	標準設定方法的選擇與執行	➢ Bookmark法理論與評量目標、試題特性，以及執行的的適切程度。
	標準設定成員的選擇與訓練	➢ 標準設定成員的多元性、專業性與代表性。 ➢ 標準設定成員對標準設定方法及流程的了解程度。
	標準設定成員的訊息回饋	➢ 第一輪提供OIB，包括各次能力及表現等級描述、試題難度、試題內容、計分原則、答案示例之評分標準、預試反應分佈曲線圖及得分人數統計。 ➢ 第一輪提供包括事實、影響及常模等回饋訊息。
	標準設定成員判斷結果的聚斂程度	➢ 檢視二輪標準設定結果的標準差是否逐漸收斂。
內部效度	標準設定成員內的一致性	➢ 檢視第一輪及第二輪決斷分數中位數差值之樣本平均數的標準誤變化。 ➢ 標準誤以0.12為檢核標準。
	標準設定成員間的一致性	➢ 檢視不同背景成員在決斷分數設定的一致性程度。 ➢ 監控決斷分數極端值。
	標準設定方法內的一致性	➢ 檢視第一輪及第二輪決斷分數中位數之樣本平均數的標準誤變化。 ➢ 標準誤以0.12為檢核標準。
外部效度	不同標準設定法的驗證	➢ 採用統計程序之群聚分析進行標準設定，作為Bookmark法所得決斷分數之輻合效度的證據。 ➢ 將Bookmark及群聚分析標準設定法的等級分類結果，進行Spearman等級相關檢定，檢核二種標準設定方法間的一致性程度。
	採用區別分析檢核Bookmark標準設定法等級分類的正確性	➢ 採用區別分析檢核Bookmark 之標準設定法，將科學探究能力分為待加強、基礎、精熟三個表現等級的正確性暨計算分類的命中率。

效度的過程證據，主要針對標準設定方法的選擇與執行、標準設定成員的選擇與訓練、標準設定成員的回饋訊息，以及標準設定成員判斷結果的聚斂程度，藉以檢核其適切性與嚴謹性；另在效度的內部證據，則依次檢核決斷分數中位數之樣本平均數的標準誤變化與獨立樣本 *t* 檢定，前者係依據學者建議（Sireci, Hauger, Wells, Shea, & Zenisky, 2009），將決斷分數中位數之樣本平均數的標準誤（standard error of the mean, $\frac{\sigma}{\sqrt{N}}=\sigma_{\bar{x}}$）變化，以不超過該評量之測量標準誤的四分之一（0.12），作為標準設定成員內判斷表現標準暨標準設定方法內之一致性的指標。後者則是透過獨立樣本 *t* 檢定所得出的平均數差值與信賴區間，進行成員間判斷表現標準之變異是否在合理範圍的檢驗，藉以比較成員間所設定的決斷分數是否存在差異；最後在效度的外部證據，有鑑於本研究係以專家判斷進行標準設定，故而本研究將針對相同資料以群聚分析之統計程序進行標準設定，找出決斷分數的結果，並比較與 Bookmark 法設定所得出之決斷分數的一致性程度，作為 Bookmark 法所得決斷分數之輻合效度的證據。同時以區別分析（discriminant analysis）檢視以試題為中心之 Bookmark 法標準設定，將科學探究能力表現分為待加強、基礎、精熟三個表現等級的正確性。

結果

一、能力評量之信度與效度檢驗

本評量之內部一致性分析結果顯示 Cronbach's α 為 .84，顯示信度良好。另以最大概似法（maximum likelihood method, ML）進行 CFA，並依據 Kline（2015）建議模式適配的統計指標包括：（1）卡方值（χ^2）顯著性；（2）RMSEA（Root Mean Square Error of Approximation）其 90%的信賴區間；（3）CFI（Comparative Fit Index）；（4）SRMR（Standardized Root Mean Square Residual），進行模式適配的主要依據，藉以考驗本評量之理論概念模式是否能為實徵資料所驗證支持。分析結果卡方分配達顯著水準（$\chi^2 = 1514.11$，$p < .001$），表示假設模式與觀察值存在顯著差異。絕對適配指標 SRMR（0.06 ≤ .06）及 RMSEA（0.06 ≤ .08）則在理想範圍；另增值適配指標 CFI 則未盡理想。因此，雖然構念效度仍有一些適配度指標未達理想，但整體而言尚在可接受範圍。

二、Bookmark 法標準設定

（一）第一輪標準設定結果

本研究將各成員在第一輪標準設定針對「觀察與定題」、「計畫與執行」、「分析與發現」、「推理與論證」等次能力，判斷最低能力受試者 67%答對率之試題並加以標籤設定的結果（表 6），進行能力估計暨轉換為 67%答對率時的能力值，求得平均值後再轉換成答對反應數及決斷分數如表 7 所示。結果顯示，「觀察與定題」、「計畫與執行」、「分析與發現」、「推理與論證」之基礎反應數依序為 7、7、5、5，精熟答對反應數則為 14、16、12、11，再將基礎及精熟答對反應數對照原始分數，即得出決斷分數。

（二）第二輪標準設定結果

本研究將各成員在第二輪標準設定針對「觀察與定題」、「計畫與執行」、「分析與發現」、「推理與論證」等次能力，判斷最低能力受試者 67%答對率之試題並加以標籤設定的結果（表 8），進行能力估計暨轉換為 67%答對率時的能力值，求得平均值後再轉換成答對反應數及決斷分數如表 7 所示。結果顯示，「觀察與定題」、「計畫與執行」、「分析與發現」、「推理與論證」之基礎答對反應數依序為 5、6、4、4，精熟答對反應數則為 12、15、12、9，再將基礎及精熟答對反應數對照原始分數，即得出決斷分數。

表 6　第一輪標準設定結果

成員	觀察與定題（Q）				計劃與執行（E）				分析與發現（A）				推理與論證（R）			
	基礎	能力	精熟	能力	基礎	能力	精熟	能力	基礎	能力	精熟	能力	基礎	能力	精熟	能力
1	6	0.02	12	1.56	7	0.37	15	1.06	5	0.12	12	1.43	3	0.11	9	1.14
2	8	0.13	13	2.06	7	0.37	17	1.24	5	0.12	12	1.43	4	0.39	11	1.73
3	9	0.14	14	2.21	9	0.65	16	1.06	4	-0.09	10	0.99	7	0.76	11	1.73
4	3	-0.32	13	2.06	7	0.37	19	2.19	5	0.12	11	1.10	6	0.72	11	1.73
5	7	0.10	15	2.29	4	0.07	11	0.83	7	0.25	12	1.43	3	0.11	9	1.14
6	6	0.02	11	0.30	6	0.12	16	1.06	5	0.12	13	1.96	5	0.43	11	1.73
7	6	0.02	14	2.21	7	0.37	13	0.98	4	-0.09	8	0.66	5	0.43	10	1.48
8	7	0.10	15	2.29	6	0.12	15	1.06	4	-0.09	10	0.99	4	0.39	10	1.48
9	4	-0.10	16	2.68	8	0.39	19	2.19	8	0.64	14	1.43	6	0.72	12	2.18
10	7	0.02	13	2.06	8	0.39	16	1.06	4	-0.09	8	0.66	2	-0.02	12	2.18
11	12	1.56	15	2.29	15	1.06	19	2.19	9	0.66	13	1.99	9	1.14	12	2.18
12	11	0.30	14	2.21	7	0.37	17	1.24	2	-1.43	7	0.25	1	-0.48	9	1.14
13	6	0.02	12	1.56	7	0.37	15	1.06	4	-0.09	12	1.43	6	0.72	11	1.73
14	6	0.02	12	1.56	8	0.39	12	0.87	4	-0.09	9	0.66	4	0.39	8	1.00
平均數	7.00	0.15	14.00	1.95	8.00	0.39	16.00	1.29	5.00	0.00	11.00	1.17	5.00	0.42	10.00	1.61
中位數	7.00	0.00	14.00	2.00	7.00	0.00	16.00	1.00	5.00	0.00	12.00	1.00	5.00	0.00	11.00	2.00
眾　數	6.00	0.02	12.00	1.56	7.00	0.37	15.00	1.06	4.00	-0.09	12.00	1.43	4.00	0.39	11.00	1.73
標準差	2.42	0.43	1.45	0.57	2.44	0.24	2.49	0.50	1.84	0.49	2.15	0.51	2.10	0.40	1.28	0.40

註：能力係為 67%答對率時的能力估計值

表 7　科學探究能力各表現等級二輪標籤設定結果

次能力	表現等級	設定標籤（第一輪）	設定標籤（第二輪）	答對反應數（第一輪）	答對反應數（第二輪）	決斷分數
觀察與定題	待加強			0-06	0-04	
	基　礎	7	5	7-13	5-11	5
	精　熟	14	12	14-17	12-17	12
計畫與執行	待加強			0-06	0-05	
	基　礎	7	6	7-15	6-14	6
	精　熟	16	15	16-20	15-20	15
分析與發現	待加強			0-04	0-03	
	基　礎	5	4	5-11	4-11	6
	精　熟	12	12	12-14	12-14	11
推理與論證	待加強			0-04	0-03	
	基　礎	5	4	5-10	4-08	4
	精　熟	11	9	11-12	9-12	9

三、標準設定的效度評估

（一）效度的過程證據

1. 標準設定方法的選擇與執行

為避免成員進行標準設定的任務負擔，本研究結合 IRT 技術，採用適用於建構與選擇題型兼具之評量卷的 Bookmark 標準設定法。

2. 標準設定成員的選擇與訓練

本研究邀請物理、化學、生物、地球科學及評量等領域之六位領域專家學者，以及包括自然科領域教師、自然科領域教科書編審教師、自然科領域學科中心教師和具備自然科領域教學經驗行政人員共八位高中自然科領域教師，成立標準設定小組，以服膺標準設定成員的組成應符合多元性、專業性與代表性的原則。同時接受本研究標準設定成員訓練，包括了解標準設定的目的與進程、標準設定方法的理論與執行說明、各表現等級之標準表現描述、標準卷的評量架構，以及標準設定的會議流程。

3. 標準設定成員的訊息回饋

本研究在第二輪標準設定進行之前，提供標準設定成員包括事實、影響及常模等回饋訊息，如此可凝聚成員的共識暨提升成員間判定的一致性。

4. 標準設定成員判斷結果的聚斂程度

本研究透過檢視二輪標準設定結果的標準差是否逐漸收斂，作為過程的效度證據之一。如表 6、表 8 所示，「觀察與定題」、「計畫與執行」、「分析與與發現」、「推理與論證」在基礎標籤設定結果之標準差變化，依序為 0.43 →0.16、0.24 →0.41、0.49 →0.46、0.40 →0.29，另在精熟設定結果之標準差變化，依序為 0.57 →0.94、0.50 →0.35、0.51 →0.37、0.40 →0.37，除了「計畫與執行」之基礎標籤及「觀察與定題」之精熟標籤外，其餘均呈現收斂趨勢，表示成員在標準設定過程當中有逐漸凝聚共識，達成表現標準的一致性。

（二）效度的內部證據

1. 標準設定成員內的一致性

本研究計算第一輪及第二輪決斷分數中位數差值之樣本平均數的標準誤變化作為標準設定成員內一致性的指標。如表 9 所示，四個次能力在第一輪到第二輪之各表現等級的標準誤，除了「觀察與定題」及「分析與發現」的精熟等級以外，其餘均在 0.12 以下，尚在可接受範圍內，表示成員內標準設定結果有逐漸凝聚共識，暨表現標準的決斷分數漸趨於一致。

表 8　第二輪標準設定結果

成員	觀察與定題（Q）				計劃與執行（E）				分析與發現（A）				推理與論證（R）			
	基礎	能力	精熟	能力	基礎	能力	精熟	能力	基礎	能力	精熟	能力	基礎	能力	精熟	能力
1	5	-0.03	11	0.30	5	0.11	13	0.98	4	-0.09	11	1.10	4	0.39	9	1.14
2	8	0.13	13	2.06	7	0.37	17	1.24	5	0.12	12	1.43	4	0.39	11	1.73
3	5	-0.03	10	0.15	6	0.12	14	1.01	8	0.64	12	1.43	3	0.11	9	1.14
4	3	-0.32	14	2.21	7	0.37	18	1.82	4	-0.09	9	0.66	4	0.39	9	1.14
5	7	0.10	12	1.56	5	0.11	15	1.06	4	-0.09	12	1.43	4	0.39	9	1.14
6	4	-0.10	10	0.15	3	-0.12	10	0.73	4	-0.09	8	0.64	2	-0.02	5	0.43
7	6	0.02	12	1.56	4	0.07	10	0.73	4	-0.09	8	0.64	4	0.39	8	1.00
8	4	-0.10	15	2.29	5	0.11	17	1.24	3	-0.41	12	1.43	3	0.11	11	1.73
9	5	-0.03	16	2.68	8	0.39	18	1.82	7	0.25	12	1.43	6	0.72	11	1.73
10	6	0.02	11	0.30	6	0.12	15	1.06	4	-0.09	9	0.66	4	0.39	9	1.14
11	7	0.10	13	2.06	7	0.37	15	1.06	7	0.25	12	1.43	5	0.43	11	1.73
12	2	-0.44	10	0.15	1	-1.21	10	0.73	2	-1.43	11	1.10	1	-0.48	9	1.14
13	5	-0.03	13	2.06	4	0.07	12	0.87	5	0.12	12	1.43	5	0.43	9	1.14
14	6	0.02	12	1.56	7	0.37	15	1.06	4	-0.09	9	0.66	5	0.43	8	1.00
平均數	5.00	-0.05	12.00	1.36	5.00	0.09	14.00	1.10	5.00	-0.08	11.00	1.11	4.00	0.29	9.00	1.24
中位數	5.00	0.00	12.00	2.00	6.00	0.00	15.00	1.00	4.00	0.00	12.00	1.00	4.00	0.00	9.00	1.00
眾　數	5.00	-0.03	13.00	2.06	7.00	0.37	15.00	1.06	4.00	-0.09	12.00	1.43	4.00	0.39	9.00	1.14
標準差	1.63	0.16	1.86	0.94	1.91	0.41	2.86	0.35	1.65	0.46	1.65	0.37	1.29	0.29	1.61	0.37

註：能力係為 67%答對率時的能力估計值

表 9　科學探究能力各表現等級二輪決斷分數中位數差值之標準誤（$N = 14$）

次能力	表現等級	第一輪－第二輪
觀察與定題	基　礎	0.11
	精　熟	0.21
計畫與執行	基　礎	0.11
	精　熟	0.09
分析與發現	基　礎	0.05
	精　熟	0.14
推理與論證	基　礎	0.09
	精　熟	0.12

1. 標準設定成員間的一致性

本研究首先將針對不同背景（領域專家 vs.領域教師、性別）成員之標準設定結果，分別就四個次能力進行平均數檢定，所得平均數差值與信賴區間如表 10 所示，透過檢視不同群組在二輪標準設定的決斷分數是否存在差異，作為標準設定成員間一致性的指標。此外，為避免成員在表現標準的判定發生極端值，進而影響決斷分數的結果，本研究將以各輪標籤設定之平均值加減兩個標準差，作為極端值界定標準，監控決斷分數極端值的發生。

（1）領域專家 vs.領域教師

不同職稱之標準設定成員其決斷分數 t 考驗的分析結果顯示，專家與教師在二輪四個次能力各表現等級標籤的信賴區間均包含零，意味不同職稱成員間所設定的決斷分數未達顯著差異。

（2）領域專家 vs.領域教師

不同職稱之標準設定成員其決斷分數 t 考驗的分析結果顯示，專家與教師在二輪四個次能力各表現等級標籤的信賴區間均包含零，意味不同職稱成員間所設定的決斷分數未達顯著差異。

（3）男成員 vs.女成員

將各成員所設定之決斷分數分成男、女二組，進行獨立樣本 t 考驗。分析結果顯示，男生組與女生組在二輪四個次能力各表現等級標籤的信賴區間均包含零，意味不同性別成員間所設定的決斷分數未達顯著差異。

（4）監控決斷分數極端值

如表 11 所示，本研究根據極端值的判定結果，檢核十四位標準設定成員在二輪四個次能力、兩個標籤設定共計 224 個設定值，結果發現共有八個判斷極端值，包括第一輪「觀察與定題」一個、「計畫與執行」一個、「分析與發現」一個；第二輪「觀察與定題」一個、「計畫與執行」三個、「分析與發現」一個。因此，本研究標準設定之極端值僅屬少數，對整體決斷分數的影響有限，同時在進行決斷分數分析時，也會將該極端值予以剔除。

2. 標準設定方法內的一致性

本研究依序計算第一輪及第二輪決斷分數中位數之樣本平均數的標準誤變化，作為標準設定方法內一致性的指標。如表 12 所示，第一輪標準設定在各次能力之基礎與精熟等級中位數的標準誤，除了「觀察與定題」和「計畫與執行」第一輪的精熟等級，以及「分析與發現」第一輪之基礎和精熟等級以外，其餘均在 0.12 以下；第二輪則除了「觀察與定題」的精熟等級之中位數標準誤在 0.12 以上，其餘均在 0.12 以下，表示決斷分數變異程度係在合理範圍，亦即標準設定結果相當一致。

表 12　科學探究能力各表現等級二輪決斷分數中位數之標準誤（*N* = 14）

次能力	表現等級	第一輪	第二輪
觀察與定題	基　礎	0.11	0.04
	精　熟	0.15	0.25
計畫與執行	基　礎	0.06	0.11
	精　熟	0.13	0.09
分析與發現	基　礎	0.13	0.12
	精　熟	0.14	0.10
推理與論證	基　礎	0.11	0.08
	精　熟	0.11	0.10

（一）效度的外部證據

有鑑於 108 課綱在新高中課程架構，所規劃的「自然科學探究與實作」課程尚未實施。是以蒐集學生關於在校科學探究與實作能力表現成績，以作為外在效標即不可行。因此，本研究另採統計程序之群聚分析進行標準設定，透過探討 Bookmark 法所得決斷分數之輻合效度，進行外在推論，藉以瞭解

Bookmark 法所得出之決斷分數的適切程度。基此，首先以群聚分析之統計程序進行標準設定，找出決斷分數的結果，並與本研究採用 Bookmark 標準設定法所得出的決斷分數相互比較一致性程度；其次採用區別分析檢核 Bookmark 之標準設定法，將科學探究能力分為待加強、基礎、精熟三個表現等級的一致性暨計算分類的命中率，作為標準設定的外部效度證據。

1. 群聚分析之標準設定

本研究以非階層性群聚法之 K 組平均法，將刪除遺漏值後共計 586 位學生分成三組，找出組內變異最小、組間變異最大的分類結果。表 13 係為各表現等級之平均數、標準差，以及答對反應數的區間範圍。本研究依據群聚分析所得決斷分數，將學生分為待加強、基礎、精熟三組，並進行變異數分析。結果顯示「觀察與定題」：F（2，583）= 1487.45，p < .001，η^2 = .84，power = 1.00；「計畫與執行」：F（2，583）= 1918.68，p < .001，η^2 = .87，power = 1.00；「分析與發現」：F（2，583）= 2154.92，p < .001，η^2 = .88，power = 1.00；「推理與論證」：F（2，583）= 1696.10，p < .001，η^2 = .85，power = 1.00，其組間變異均達顯著水準，效果量（η^2）均屬高度關聯強度，統計考驗力均為 1.00，分析推論犯第二類型錯誤機率（Type Ⅱ error）0.00%，分類一致性相當高。

2. 不同標準設定法的驗證

本研究首先以群聚分析之 k 組平均法進行標準設定法，作為外在效標，藉以檢核 Bookmark 標準設定法的外部效度。其次，依據這二種標準設定方法所得出的決斷分數將學生分成待加強、基礎、精熟三個表現等級進行 Spearman 等級相關檢定，藉以求出二種標準設定方法間的關聯程度。分析結果顯示，根據二種標準設定法所得出的決斷分數，分別將學生在「觀察與定題」、「計畫與執行」、「分析與發現」及「推理與論證」分為三個表現等級的相關係數依序為 .66（p < .01）、.78（p < .01）、.89（p < .01）、.91（p < .01），均達顯著水準，表示二種標準設定法在判斷表現等級有相當程度的一致性。

表 10　不同背景成員在二輪次能力各表現等級決斷分數 *t* 檢定

背景變項	輪次	類別	人數	觀察與定題		計畫與執行		分析與發現		推理與論證	
				基礎	精熟	基礎	精熟	基礎	精熟	基礎	精熟
職稱別	一	領域專家	6	-0.23	0.10	-0.28	-0.15	0.12	0.00	-0.00	-0.35
		領域教師	8	(-0.73~0.28)	(-0.21~0.42)	(-0.33~0.27)	(-0.75~0.46)	(-0.47~0.71)	(-0.62~0.62)	(-0.50~0.49	(-0.79~0.91)
	二	領域專家	6	0.05	-0.10	0.27	0.17	-0.00	0.02	0.10	-0.04
		領域教師	8	(-0.14~0.24)	(-1.25~1.05)	(-0.20~0.73)	(-0.25~0.58)	(-0.65~0.65)	(-0.43~0.47)	(0.21~0.42	(-0.50~0.42)
性別	一	男	4	-0.30	0.17	-0.16	-0.24	-0.59	-0.29	0.21	-0.18
		女	10	(-0.84~0.25)	(-0.20~0.53)	(-0.34~0.31)	(-0.69~0.65)	(-0.71~0.59)	(-0.94~0.36)	(-0.31~0.73)	(-0.71~0.35)
	二	男	4	-0.04	0.68	0.18	0.02	-0.24	-0.36	0.17	-0.24
		女	10	(-0.25~0.17)	(-0.51~1.87)	(-0.35~0.72)	(-0.44~0.49)	(-0.93~0.46)	(-0.80~0.79)	(0.20~0.53)	(-0.71~0.24)

表 11　二輪次能力各表現等級極端值範圍

輪次	標籤設定	觀察與定題				計畫與執行				分析與發現				推理與論證			
		基礎		精熟		基礎		精熟		基礎		精熟		基礎		精熟	
		反應數	能力值	反應數	能力值	反應數	能力值	反應數	能力值	反應數	能力值	反應數	能力值	反應數	能力值	反應數	能力值
第一輪	平均數	7	0.15	14	1.95	8	0.39	16	1.29	5	0.00	11	1.17	5	0.42	10	1.61
	M-2*SD	2	-0.71	11	0.81	3	-0.09	11	0.29	1	-0.98	6	0.15	0	-0.38	8	0.81
	M+2*SD	12	1.01	16	3.09	12	0.87	21	2.29	9	0.98	15	2.19	9	1.22	13	2.41
第二輪	平均數	5	-0.05	12	1.36	5	0.09	14	1.10	5	-0.08	11	1.11	4	0.29	9	1.24
	M-2*SD	2	-0.37	8	-0.52	1	-0.73	8	0.40	2	-1.00	8	0.37	1	-0.29	6	0.50
	M+2*SD	8	0.27	16	3.24	9	0.91	20	1.80	8	0.84	14	1.85	7	0.87	12	1.98

表 13　群聚分析敘述統計摘要表

能力向度	表現等級	人數	平均數	標準差	標準誤	最小值	最大值	決斷分數
觀察與定題	待加強	146	4.88	1.19	.10	0	6	
	基　礎	224	8.01	.83	.06	7	9	7
	精　熟	216	11.00	1.15	.08	10	16	10
	合　計	586	8.33	2.59	.11	0	16	
計畫與執行	待加強	167	4.09	2.11	.16	0	7	
	基　礎	215	10.20	1.35	.09	8	12	8
	精　熟	204	14.87	1.57	.11	13	19	13
	合　計	586	10.09	4.59	.19	0	19	
分析與發現	待加強	147	3.42	1.25	.10	0	5	
	基　礎	188	7.69	1.10	.08	6	9	6
	精　熟	251	11.21	1.12	.07	10	14	10
	合　計	586	8.13	3.32	.14	0	14	
推理與論證	待加強	131	2.05	.96	.08	0	3	
	基　礎	273	5.45	1.10	.07	4	7	4
	精　熟	182	9.24	1.16	.09	8	12	8
	合　計	586	5.87	2.84	.12	0	12	

1. 區別分析（discriminant analysis）

本研究以科學探究能力各次能力的總分為預測變項（predictor variable），以 Bookmark 標準設定法得出的決斷分數所分類出待加強、基礎、精熟等表現等級為效標變項（criterion variable），進行區別分析，亦即透過各次能力的總分組合成一個最有效的分類函數，建立區別函數，藉以檢視 Bookmark 標準設定法將科學探究能力分為三個表現等級的一致性。

分析結果如表 14 所示，區別函數之顯著性考驗結果顯示，「觀察與定題」：F（2，583）= 593.56，$\lambda = .33$，$p < .001$；「計畫與執行」：F（22，583）= 1064.07，$\lambda = .22$，$p < .001$；「分析與發現」：F（22，583）= 1812.88，$\lambda = .14$，$p < .001$；「推理與論證」：F（2，583）= 1286.00，$\lambda = .19$，$p < .001$，其各表現等級之 F 考驗達顯著差異，表示四個次能力的總分對於學生在表現等級的分類上有顯著的預測力，亦即可以有效區別學生在各次能力之待加強、基礎、精熟三個表現等級。由交叉分析結果顯示（表 15），「觀察與定題」、「計畫與執行」、「分析與發現」，及「推理與論證」的整體分類一致性，依序為 79.50%、86.00%、100.00%、89.90%。各次能力在待加強及精熟二個表現等級的一致性預測率均為 100.00%，而基礎之表現等級的一致性預測率亦達七至八成以上（觀察與定題：72.60%、計畫與執行：77.53%、分析與發現：100.00%，推理與論證：82.23%）。

綜上所述，Bookmark 法與群聚分析二種標準設定法在判斷表現等級之輻合效度，除「觀察與定題」呈現中度相關以外，其他如「計畫與執行」、「分析與發現」及「推理與論證」均呈高度相關。另區別分析結果顯示，Bookmark 法將各次能力分為待加強與精熟表現等級的一致性均達 100.00%，另「分析與發現」分為基礎等級的一致性亦達 100.00%，其他如「觀察與定題」、「計畫與執行」及「推理與論證」分為基礎等級的一致性也有七至八成的一致性，表示 Bookmark 標準設定法將科學探究能力區分為三個表現等級效果良好。

表 14　科學探究能力在各表現等級之區別分析摘要表

	標準化典型區別係數	結構係數	未標準化典型區別係數	截距	區別函數（λ）	Wilks' ∧	χ^2
觀察與定題	1.00	1.00	.67	-5.59	2.04	.33	647.49***
計畫與執行	1.00	1.00	.47	-4.73	3.65	.22	896.03***
分析與發現	1.00	1.00	.81	6.57	6.22	.14	1152.44***
推理與論證	1.00	1.00	.82	-4.80	4.41	.19	984.43***

註1：典型相關係數係指區別分數與組別間的關聯程度，相當於變異數分析中的效果量（η）。
註2：Wilk's lambda（∧）係為組內離均差平方和與總離均差平方和的比（SS_w/SS_t）
***$p < .001$.

表 15　科學探究能力各表現等級之分類一致性叉表

能力向度	表現等級	決斷分數	實際分類人數	預測結果分類（人數／百分比）					
				待加強（%）		基礎（%）		精熟（%）	
觀察與定題	待加強		91	91	（100.00）	0	（0.00）	0	（0.00）
	基　礎	5	438	55	（12.56）	318	（72.60）	65	（14.84）
	精　熟	12	57	0	（0.00）	0	（0.00）	57	（100.00）
	總預測一致性			79.50%					
計劃與執行	待加強		116	116	（100.00）	0	（0.00）	0	（0.00）
	基　礎	6	365	28	（7.67）	283	（77.53）	54	（14.79）
	精　熟	15	105	0	（0.00）	0	（0.00）	105	（100.00）
	總預測一致性			86.00%					
分析與發現	待加強		147	147	（100.00）	0	（0.00）	0	（0.00）
	基　礎	6	273	0	（0.00）	273	（100.00）	0	（0.00）
	精　熟	11	166	0	（0.00）	0	（0.00）	166	（100.00）
	總預測一致性			100.00%					
推理與論證	待加強		131	131	（100.00）	0	（0.00）	0	（0.00）
	基　礎	4	332	0	（0.00）	273	（82.23）	59	（17.77）
	精　熟	9	123	0	（0.00）	0	（0.00）	123	（100.00）
	總預測一致性			89.90%					

討論

一、綜合討論

「科學探究能力」標準設定的任務，係將「觀察與定題」、「計畫與執行」、「分析與發現」及「推理與論證」四個次能力，分為待加強、基礎及精熟三個表現等級。茲將盧列研究結果暨綜合討論，並針對本研究結果，提出未來可供學術社群應用，暨研究者仍待深究議題的建議。

（一）科學探究能力標準設定可獲得過程效度證據的支持

本研究首先考量適用多題型評量的優勢暨成員進行標準設定的任務負擔，採用執行容易的 Bookmark 標準設定法。

其次，成員選擇除涵蓋物理、化學、生物、地球科學及評量領域的學者專家外，也包括參與教科書編審、學科中心及行政工作等實務教學經驗的高中教師，因此標準設定成員檢具多元性、專業性與代表性。這些成員均必須接受訓練，以了解包括標準設定的主旨目的、理論概念和執行步驟，以及熟悉各表現等級的標準表現描述、標準卷的評量架構與標準設定會議的流程。

再者，由於每個成員判斷最低能力受試者在各表現等級 67%答對率的試題，是一項相當具有挑戰的認知任務，這中間涉及到二個重要的判斷關鍵，其一為最低能力受試者，其二為 67%答對率。是以，成員為掌握這二者頗具抽象的關鍵判斷準則暨順利完成設定任務，往往會參照教學班

級學生的反應，透過了解教學班級學生程度的優勢，推論可能的答題表現，以利基礎和精熟的標籤設定。然則不同學校的學生表現係存在相當程度的差異，故而成員在參照其答題反應時，即有主觀判斷的疑慮。為使成員對於各次能力在每個表現等級之標準表現的認知與判斷能夠漸趨一致，本研究提供事實、影響及常模等的回饋訊息，以利凝聚成員對於各表現等級之標準判定的共識。

最後，分析成員判斷結果之聚斂程度發現，二輪標準設定結果的標準差，除了「計畫與執行」的基礎標籤及「觀察與定題」的精熟標籤外，其他設定結果均呈現收斂趨勢，表示成員在設定過程對於各等級標準表現的共識漸趨一致。綜上所述，本研究科學探究能力標準設定可獲得過程效度證據的支持。

（二）科學探究能力標準設定可獲得內部效度證據的支持

本研究首先在二輪決斷分數中位數差值之樣本平均數的標準誤（SE）估計結果顯示，第一輪標準設定在各次能力之基礎及精熟等級中位數差值的標準誤，除了「觀察與定題」（0.21）與「分析與發現」（0.14）之精熟等級超過評量之測量標準誤的四分之一（> 0.12），其餘均介於 0.05- 0.11（< 0.12），表示成員內標準設定結果檢具一致性。

其次，二輪決斷分數中位數之樣本平均數的標準誤（SE）估計結果發現，第一輪除「觀察與定題」（0.15）、「計畫與執行」（0.13）、「分析與發現」的基礎（0.13）和精熟（0.14）等級，以及第二輪「觀察與定題」之精熟等級（0.25）超過評量之測量標準誤的四分之一（> 0.12），其餘介於 0.06- 0.11（< 0.12），表示決斷分數變異程度係在合理範圍，亦即標準設定結果相當一致。

再者，本研究依據職稱別（領域專家 vs.領域教師）及性別等變項將成員分組，進行獨立樣本 *t* 檢定，分析結果顯示不同背景變項之標準設定成員在二輪四個次能力各表現等級所設定的決斷分數均未達顯著差異，具有良好的一致性。另標準設定極端值的監控結果發現，僅有少數極端值出現，故而對於整體決斷分數的影響甚微，同時在進行決斷分數分析時，也會將該極端值予以剔除，以期更加優質及精確標準設定結果，因此標準設定成員間有相當程度的一致性。綜上所述，本研究科學探究能力標準設定可獲得內部效度證據的支持。

（三）科學探究能力標準設定可獲得外部效度證據的支持

本研究以群聚分析之 K 組平均法與 Bookmark 法二種標準設定所得到的決斷分數，進行 Spearman 等級相關檢定。分析結果顯示，二種標準設定結果在「觀察與定題」、「計畫與執行」、「分析與發現」及「推理與論證」分為三個表現等級的相關係數均達顯著水準，表示在判斷表現等級有相當程度的一致性。基此，本研究採用群聚分析的分類結果，作為探討 Bookmark 法有效性之輻合證據可獲得支持。

其次，分別以 Bookmark 法得出之決斷分數所分出待加強、基礎、精熟表現等級為效標變項、四個次能力總分為預測變項，進行區別分析。結果顯示，Bookmark 法在「觀察與定題」、「計畫與執行」、「分析與發現」及「推理與論證」整體分類一致性依序為 79.50%、86.00%、100.00%、89.90%，另四個次能力在待加強、精熟等級，及「分析與發現」在基礎等級分類的一致性均達 100.00%；「觀察與定題」（72.60%）、「計畫與執行」（77.50%）及「推理與論證」（82.20%）於基礎等級分類的一致性達七至八成，足見 Bookmark 標準設定法所得出的決斷分數在各表現等級分類之區別力相當高。因此，本研究科學探究能力標準設定可獲得外部效度證據的支持。

綜上所述，本研究採用多媒體電腦化評量進行能力取向之評測，有別於過去研究所使用的紙本測驗，加上進行標準設定時，除提供成員關於評量之紙本資料外，另必須在現場準備電腦讓成員能夠觀測學生實際所看到的試題介面，以利其判斷。研究結果顯示，在提供成員多媒體電腦化評量動畫視窗的前提下，以 Bookmark 法進行多媒體電腦化之科學探究能力評量的標準設定，係獲得過程、內部及外部等多元效度證據的支持。

（四）Bookmark 與群聚分析標準設定方法的比較

依據本研究二輪 Bookmark 法所得出之決斷分數，將 605 位高中生進行等級分類，結果發現次能力之各表現等級的通過百分比如表 16 所示，待加強、基礎、精熟通過比率範圍依序為 64%~90%、8%~49%、0.7%~7.4%，其中半數以上均落在待加強，另基礎等級則以「推理與論證」佔有 8 成最高、「計畫與執行」只佔 1 成左右最低，至於落在精熟比例則最小，不到 1 成。比較兩種標準設定所得出的決斷分數發現，Bookmark 法在基礎等級之決斷分數除了「觀察與定題」及「計畫與執行」

低於群聚分析法以外，其餘均相同；四個子能力在精熟等級之決斷分數均為 Bookmark 法高於群聚分析法（參照表 13、表 15）。因此，通過 Bookmark 法所得到的標準較群聚分析嚴格，顯示本研究標準設定結果據以分類時似有標準偏高的現象。此現象與 Green 等人（2003）、Reckase（2006）認為 Bookmark 法所得出決斷分數易出現負向偏誤的結論並不一致。而造成此結果的原因之一，可能是標準設定成員係以學生接受「自然科學探究與實作」課程後的表現做為標準判斷的依據；然而，108 課綱尚未上路，本研究這些受測學生並未真正接受科學探究與實作的課程，致使評測結果大都落在待加強等級並不意外，但未來有必要在課綱實施後，蒐集已接受科學探究與實作課程學生的資料，檢視其在各表現等級的通過率，檢核 Bookmark 法的嚴格程度。

表 16　科學探究能力各表現等級通過百分比與人數

能力向度		實際資料通過百分比（人）		
		待加強	基礎	精熟
觀察與定題（Q）	第二輪	64.3%（389）	34.5%（209）	1.2%（07）
	第一輪	64.3%（389）	35.5%（215）	0.2%（01）
計畫與執行（E）	第二輪	84.8%（513）	14.5%（088）	0.7%（04）
	第一輪	89.9%（544）	09.4%（057）	0.7%（04）
分析與發現（A）	第二輪	26.4%（160）	66.1%（400）	7.4%（45）
	第一輪	43.5%（263）	49.1%（297）	7.4%（45）
推理與論證（R）	第二輪	90.9%（550）	08.1%（049）	0.1%（06）
	第一輪	90.9%（550）	08.1%（049）	0.1%（06）

二、研究建議

本研究建議未來高中正式實施「自然科學探究與實作」課程後，可植基科學探究能力標準設定的結果，繼續蒐集真正接受「探究與實作」課程學生的評測反應，延續進行標準設定的研究與效度驗證，包括各次能力在各表現等級之標準表現的界定（PLDs）、學生科學探究與實作成績作為效標之外部效度評估證據，以及開發科學探究能力評量報表系統。

（一）持續修訂各次能力 PLDs

PLDs 係為成員進行 Bookmark 標準設定時的重要參酌資料，其各次能力的表現描述與各表現等級的判斷準則，應當盡可能具體暨獲設定成員一致認同的共識。本研究在二次標準設定過程中，均觀察到成員對於各次能力及表現等級的描述仍存有疑義，尤其是高中教師，在會議中多次提出討論，意味著本研究所發展之 PLDs 仍與部分成員的觀點有所出入，暨在實務教學應用上可能未盡理想。然則 108 課綱尚未正式實施，是以當新高中課程架構所規劃的「自然科學探究與實作」課程啟動開跑後，會呈現如何的態樣跟局面，有待進一步觀察和資料蒐集，從而評估修訂增補 PLDs 的可能性，使其能夠更加貼近教學場域的需求既符合新課綱的課程主旨。

（二）持續蒐集外部效度證據

由於受限 108 課綱所規劃之新高中「自然科學探究與實作」課程尚未正式實施，因此造成本研究蒐集受試者能力資料的困難，致使研究者採用不同標準設定方法作為效標進行外在推論，藉以作為外部效度的證據。基此，本研究建議待新課綱正式實施後，應蒐集學生「自然科學探究與實作」課程的學習表現作為效標，進行標準設定外部效度的評估，以作為複核效度證據。

（三）開發科學探究能力評量報表系統

本研究發展科學探究能力評量標準，一則可避免高中師生對評量結果詮釋和理解的困難，二則還能提供教師評定和描述高中生科學探究能力學習表現的參考，藉以檢核是否達到課程所預期的目標或程度。不僅可緊密連結課程綱要、教學與評量，亦可作為學力監控的方法。因此，本研究建議未來可依據標準設定所制訂各標準表現描述（PLDs）與決斷分數，開發評量報表系統，提

供具診斷功能之分析式的回饋評述，來協助師生了解科學探究能力的強、弱項，作為教師教學與學生加強的參酌依據。

參考文獻

吳清山（2014）：標準參照測驗。**教育資料與研究，113**，205-206。DOI：10.6724/ERR.201405_(113).0007。[Wu, C. S. (2014). Standard reference test. *Educational Resources and Research, 113*, 205-206. DOI: 10.6724/ERR.201405_(113).0007]

吳宜芳、鄒慧英、林娟如（2010）：標準設定效度驗證之探究：以大型數學學習成就評量為例。**測驗學刊，57**（1），1-27。DOI：10.7108/PT.201003.0001。[Wu, Y. F., Tzou, H., & Lin, J. R. (2010). Validating the performance standards for cut scores in a large-scale mathematics assessment. *Psychological Testing, 57*(1), 1-27. DOI: 10.7108/PT.201003.0001]

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。DOI：10.6209/JORIES.2018.63(1).06。[Lin, H. H., Lin, S. H., & Wu, H. K. (2018). Developing and validating a constructed-response assessment of Sientific abilities: A case of the optics unit. *Journal of Research in Education Sciences, 63*(1), 173-205. DOI: 10.6209/JORIES.2018.63(1).06]

陳慧娟（2015）：「師生共同增能」與「學生增能」教學實驗方案促進偏遠地區國中學生知識信念，自我調整策略與科學學習成就之比較研究。**教育科學研究期刊，60**（4），21-53。DOI：10.6209/JORIES.2015.60(4).02。[Chen, H. J. (2015). Effects of empowerment programs in remote junior high schools on scientific epistemological beliefs, self-regulation strategies, and academic achievement in science. *Journal of Research in Education Sciences, 60*(4), 21-53. DOI: 10.6209/JORIES.2015.60(4).02]

曾建銘、王暄博（2012a）：標準設定之效度評估：以 TASA 國語科為例。**教育學刊，39**，77-118。DOI：10.3966/156335272012120039003。[Cheng, C. M, & Wang, H. P. (2012a). Assessing the standards set by TASA and its standard-setting procedures. *Educational Review, 39*, 77-118. DOI: 10.3966/156335272012120039003]

曾建銘、王暄博（2012b）：臺灣學生學習成就評量資料庫標準設定探究：以 2009 年國小六年級社會科為例。**教育與心理研究，35**（3），115-149。[Cheng, C. M, & Wang, H. P. (2012b). A primary study on the standard setting of the Taiwan Assessment of Student Achievement considering 6th grade social students in 2009 as an example. *Journal of Educational and Psychology, 35*(3), 115-149.]

謝名娟、謝進昌、林世華（2013）：不同方法設定英文科決斷分數之實務性研究。**測驗學刊，60**（3），519-544。DOI：10.7108/PT.201212.0513。[Hsieh, M. C., Hsieh, J. C., & Lin, S. H (2013). The comparison of two standard setting method on an English test. *Psychological Testing, 60*(3), 519-544. DOI: 10.7108/PT.201212.0513]

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (pp. 508-600). Washington, DC: American Council on Education.

Angoff, W. H. (1984). *Scales, norms, and equivalent scores*. Princeton, NJ: Educational Testing Service.

Berk, R. A. (1986). A consumer's guide to setting performance standards on criterion-referenced tests. *Review of Educational Research, 56*(1), 137-172. DOI: 10.3102/00346543056001137

Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225-258). Mahwah, NJ: Lawrence Erlbaum Associates.

Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests.* Thousand Oaks, CA: Sage.

Ebel, R. L., & Frisbie, D. A. (1986). *Essentials of educational measurement* (4th ed.). Englewood Cliffs, NJ: Prentice-Hall.

Giraud, G., Impara, J. C., & Plake, B. S. (2005). Teachers' conceptions of the target examinee in Angoff standard setting. *Applied Measurement in Education, 18*(3), 223-232. DOI: 10.1207/s15324818ame1803_2

Green, D. R., Trimble, C. S., & Lewis, D. M. (2003). Interpreting the results of three different standard-setting procedures. *Educational Measurement: Issues and Practice, 22*(1), 22-32. DOI: 10.1111/j.1745-3992.2003.tb00113.x

Hambleton, R. K., Jaeger, R. M., Plake, B. S., & Mills, C. (2000). Setting performance standards on complex educational assessments. *Applied Psychological Measurement, 24*(4), 355-366. DOI: 10.1177/01466210022031804

Hambleton, R. K. (2001). Setting performance standards on educational assessments and criteria for evaluating the process. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 89-116). Mahwah, NJ: Lawrence Erlbaum Associates.

Hsu, Y. S., Chang, H. Y., Fang, S. C., & Wu, H. K. (2015). Developing technology-infused inquiry learning modules to promote science learning in Taiwan. In M. S. Khine (Ed.), *Science education in East Asia: Pedagogical innovations and research-informed practices* (pp. 373-403). Dordrecht: Springer International Publishing. DOI: 10.1007/978-3-319-16390-1_15

Hsu, Y. S., Wu, H. K., & Hwang, F. K. (2008). Fostering high school students' conceptual understandings about seasons: The design of a technology-enhanced learning environment. *Research in Science Education, 38*(2), 127-147. DOI: 10.1007/s11165-007-9041-1

Huynh, H. (2006). A clarification on the response probability criterion RP67 for standard settings based on bookmark and item mapping. *Educational Measurement: Issues and Practice, 25*(2), 19-20. DOI: 10.1111/j.1745-3992.2006.00053.x

Impara, J. C., & Plake, B. S. (1997). Standard setting: An alternative approach. *Journal of Educational Measurement, 34*(4), 353-366. DOI: 10.1111/j.1745-3984.1997.tb00523.x

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4*, 461-475. DOI: 10.3102/01623737004004461

Kane, M. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425-461. DOI: 10.3102/00346543064003425

Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement: Issues and Practice, 25*(1), 4-12. DOI: 10.1111/j.1745-3992.2006.00047.x

Kline, R. B. (2015). *Principles and practice of structural equation modeling, fourth edition* (4th ed.). New York: Guilford Press.

Lewis, D. M., Mitzel, H. C., & Green, D. R. (1996). *Standard setting: A bookmark approach*. Paper presented at the council of chief state school officers national conference on large scale assessment, Boulder, CO.

Linn, R. L., & Herman, J. L. (1997). *A policymaker's guide to standards-led assessment*. Denver, CO: Education Commission of the States.

Livingston, S. A., & Zieky, M. J. (1989). A comparative study of standard-setting methods. *Applied Measurement in Education, 2*(2), 121-141. DOI: 10.1207/s15324818ame0202_3

Loomis, S. C. (2000, April). *Feedback in the NAEP achievement levels setting process*. Paper presented at the meeting of the national council on measurement in education, New Orleans.

Loomis, S. C., & Bourque, M. L. (2001). From tradition to innovation: Standard setting on the national assessment of educational progress. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 175-217). Mahwah, NJ: Erlbaum.

Mitzel, H. C., Lewis, D. M., Patz, R. J., & Green, D. R. (2001). The bookmark procedure: Psychological perspectives. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 249-281). Mahwah, NJ: Lawrence Erlbaum Associates.

Nedelsky, L. (1954). Absolute grading standards for objective tests. *Educational And Psychological Measurement, 14*, 3-19.

Peterson, C. H., Schulz, E. M., & Engelhard, G. (2011). Reliability and validity of bookmark-based methods for standard setting: Comparisons to Angoff-based methods in the national assessment of educational progress. *Educational Measurement: Issues and Practice, 30*(2), 3-14. DOI: 10.1111/j.1745-3992.2011.00200.x

Raymond, M. R., & Reid, J. B. (2001). Who made thee a judge? Selecting and training participants for standard-setting. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 119-157). Mahwah, NJ: Lawrence Erlbaum Associates.

Reckase, M. D. (2000). *The evolution of the NAEP achievement levels setting process: A summary of the research and development efforts conducted by ACT.* Iowa City, IA: American College Testing.

Reckase, M. D. (2001). Innovative methods for helping standard-setting participants to perform their task: The role of feedback regarding consistency, accuracy, and impact. In G. J. Cizek (Ed.), *Setting performance standards: Concepts, methods, and perspectives* (pp. 159-173). Mahwah, NJ: Lawrance Erlbaum Associates.

Reckase, M. D. (2006). A conceptual framework for a psychometric theory for standard setting with examples of its use for evaluating the functioning of two standard setting methods. *Educational Measurement: Issues and Practice, 25*(2), 4-18. DOI: 10. 1111/j.1745-3992.2006.00052.x

Sireci, S. G., Hauger, J. B., Wells, C. S., Shea, C., & Zenisky, A. L. (2009). Evaluation of the standard setting on the 2005 Grade 12 National Assessment of Educational Progress mathematics test. *Applied Measurement in Education, 22*(4), 339-358. DOI: 10.1080/08957340903221659

Sturmberg, J. P., & Hinchy, J. (2010). Borderline competence-from a complexity perspective: Conceptualization and implementation for certifying examinations. *Journal of Evaluation in Clinical Practice, 16*(4), 867-872. DOI: 10.1111/j.1365-2753.2010.01501.x

Timm, N. H. (2002). *Applied multivariate analysis*. New York: NY: Springer-Verlag.

Violato, C., Marini, A., & Lee, C. (2003). A validity study of expert judgment procedures for setting cutoff scores on high-stakes credentialing examinations using cluster analysis. *Evaluation & the Health Professions, 26*(1), 59-72. DOI: 10.1177/0163278702250082

Wu, H. K. (2010). Modelling a complex system: Using novice-expert analysis for developing an effective technology-enhanced learning environment. *International Journal of Science Education, 32*(2), 195-219. DOI: 10.1080/09500690802478077

Wu, H. K., & Hsieh, C. E. (2006). Developing sixth graders' inquiry skills to construct scientific explanations in inquiry-based learning environments. *International Journal of Science Education, 28*(11), 1289-1313. DOI: 10.1080/09500690600621035

Wu, H. K., Kuo, C. Y., Jen, T. H., & Hsu, Y. S. (2015). What makes an item more difficult? Effects of modality and type of visual information in a computer-based assessment of scientific inquiry abilities. *Computers & Education, 85*, 35-48. DOI: 10.1016/j.compedu.2015.01.007

Yin, P., & Schulz, E. M. (2005, April). *A comparison of cut scores and cut score variability from Angoff-based and Bookmark-based procedures in standard setting*. Paper presented at the annual meeting of the national council on measurement in education, Montreal, Canda.

收 稿 日 期：2018 年 02 月 26 日
一稿修訂日期：2018 年 05 月 09 日
二稿修訂日期：2018 年 06 月 27 日
三稿修訂日期：2018 年 07 月 18 日
四稿修訂日期：2018 年 08 月 07 日
五稿修訂日期：2018 年 08 月 30 日
接受刊登日期：2018 年 08 月 30 日

Bulletin of Educational Psychology, 2019, 50(3), 473-502
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

Validating the Standard Setting on Multimedia-based Assessment of Scientific Inquiry Abilities

Hsiao-Hui Lin
College of Education
National Taiwan Normal University

Hsin-Kai Wu
Graduate Institute of Science Education
National Taiwan Normal University

This study developed a standard setting for Grade 11 of the Multimedia-based Assessment of Scientific Inquiry Abilities (MASIA) based on three levels of standard performance descriptions: below basic, basic, and proficient. The study also used a bookmark to identify the cut-off scores. Furthermore, the study discussed the correct degree of the MASIA standard, which depends on multiple levels of evidence, namely procedural, internal, and external evidence for validity. First, the result of the procedural evaluation for validity showed that the standard setting of scientific inquiry abilities adopted in this study is supported by the procedural evidence for validity. Second, the result of the internal evaluation for validity showed that the standard error of each performance level that participants reached in rounds one and two were within an acceptable range (standard error [SE] < 0.12), thus indicating good intra-rater consistency. The consistency within the standard-setting method was evaluated using the standard error of the sample mean based on the median of the cut-off scores from round two. The result showed that the standard error of every performance level was within an acceptable range (SE < 0.12), thus denoting high consistency within the results of the standard-setting method. Third, the inter-rater consistency of the standard setting was examined using an independent sample *t* test, and the results showed that none the cut-off scores set by the participants of different groups reached statistical significance. Therefore, the standard setting of scientific inquiry abilities can be supported by internal evidence for procedural validity. Finally, this study treated the quasi-setting results derived from the cluster analysis as convergent validity-based evidence to assess external validity. The results showed that the correlation coefficient of the three performance levels of the students differentiated by two standard-setting methods reached statistical significance, thus indicating that those who judged the performance levels of the students had a certain degree of consistency. Moreover, a discriminant analysis was conducted to determine the consistency of the standard settings, and the results revealed that the consistency of the factors classified into “observing and questioning,” “planning and experimenting,” “analyzing and concluding,” and “reasoning and arguing” were in the following sequence: 79.50%, 86.00%, 100.00%, and 89.90%. In the present study, the cut-off scores obtained by the bookmark presented high discrimination for each performance level category and can be supported using external evidence for validity. These results suggest that the bookmark standard setting of scientific inquiry abilities are appropriate and effective.

KEY WORDS: Bookmark, Scientific inquiry abilities, Standard setting, Validation