國立臺灣師範大學教育心理與輔導學系
教育心理學報，民 98，41 卷，1 期，69－90 頁

以常態混組模型討論書籤標準設定法對英語聽讀基本能力標準設定有效性之輻合證據*

吳毓瑩　國立台北教育大學 心理與諮商學系
陳彥名　國立台北教育大學 教育學系
張郁雯　國立台北教育大學 教育學系
陳淑惠　國立台北教育大學 兒童英語教育學系
何東憲　國立中正大學 心理學系
林俊吉　國立中正大學 心理學系

本研究旨在探討書籤標準設定法（簡稱書籤標定法）應用於 2005 年台灣學生學習成就資料庫（TASA, Taiwan Assessment of Student Achievement）中之「英語文學習成就評量」的英語聽讀基本能力標準設定（Standard Setting）的判斷歷程，以及所判斷結果之輻合證據的有效性。研究樣本有 10101 名小六學生，分層抽樣自全台各地，施測方式採取平衡不完全區塊設計（Balanced Incomplete Block Design），將 70 個聽與讀的題目設計成 6 個 40 題組成的題本。書籤標定法標準設定會議由十三位專家（七位學科內容教授、三位測驗教授、以及三位英語專家教師）組成，會中共識設一切分點（Θ=-0.57），通過組學生佔全體 72.9%。本研究利用常態混組模型（normal mixture model）之計量模式結果作爲書籤標定法有效程度的輻合證據（convergent evidence），其估計的切分點（Θ=-0.40）與專家設定的分類結果達到 .87 之 Kappa 一致性。文末研究者提出實務使用上以及理論上的討論議題。

關鍵詞：英語聽讀能力、效度之輻合證據、常態混組模型、書籤標定法、標準設定

一、研究目的與問題

我國於民國八十九年頒布九年一貫課程暫行綱要（教育部，民89），明定九十學年度開始，除了原有國中階段的英語課程之外，英語課程向下延伸至國小階段。至於所實施的英語教學成效如何，

* 此研究之完成，感謝國立教育研究院籌備處計畫項目 NAER-94-12-A-1-01-02-3-02 以及 NAER-95-12-A-1-01-02-3-02 之資助。通訊作者：吳毓瑩，通訊方式：wuyuhyin@gmail.com。

臺灣學生英語文學習成就評量（Taiwan Assessment of Student Achievement-English, TASA-EN）乃於民國九十四年五月進行全國一萬零四百五十五名國小六年級學童之抽測，並完成相關資料的建檔及分析（陳淑惠、吳毓瑩、張郁雯、何東憲，民95；陳淑惠、吳毓瑩、何東憲、張郁雯、陳錦芬，民94）。此評量乃依據九年一貫課程綱要中之英語文學習領域分段能力指標（教育部，民92），以及能力指標的解讀與示例（教育部，民93），採取標準本位評量（standard-based assessment）的精神設計測驗，以瞭解國小六年級學生學習英語一個學年後（每週兩堂課，每堂40分鐘）英語聽讀基本能力的狀況。TASA-EN 測量的內容向度與題目分佈見陳淑惠等（民95）。

本研究採用「書籤標準設定法」（bookmark standard setting method，簡稱書籤標定法），設定英語文學習成就表現標準（performance standards）的切分點，以檢視受試者是否達到能力指標所定義的表現（Reckase, 2006）。Perie 於2005年指出美國有31州採用書籤標定法作爲州級評量學生成就表現之標準設定的方法（引自 Karantonis and Sireci, 2006）。雖然書籤標定法廣泛用於表現標準之設定，但是研究成果卻相對稀少（Karantonis & Sireci, 2006）。Reckase 認爲書籤標定法統計基礎薄弱，用以設定表現標準的切分點，有潛在風險。因此，發展計量效標來檢核切分點的適當性，是重要的研究議題。

本研究目的有二，首先，探討以書籤標定法，設定英語文學習成就表現標準之可行性；其次，呼應 Reckase（2006）之提醒，本研究利用常態混組模型之模式找出潛在組別，做爲計量效標，評估書籤標定法所得出的表現標準有效性的輻合證據。

二、文獻探討

本節首先說明標準設定的起源以及標準的意義，以了解書籤標定法廣泛應用的脈絡；接著討論書籤標定法之程序以及優缺點；最後簡介常態混組模型，並說明何以其適合作爲書籤標定法的輻合證據以提供計量之效標。

（一）標準設定之起源

標準設定的由來，在測驗發展史上，可追溯至 Flanagan（1951）於第一版〈教育測量〉（Educational Measurement）一書中提出效標參照測驗的想法。七十年代時，Angoff（1971）提出數種標準設定方法，Jaeger（1989）在第三版〈教育測量〉中將各類標準設定方法系統化整理出來，方法雖異，目的卻是雷同，因當時美國正盛行「最低能力檢測」（Minimum Competency Test），故標準設定乃在設立最低能力之切分點，並篩選出通過切分點的學生。因應當時最低能力檢測的要求，研究者皆使用一個切分點的標準設定法，將受試者分爲兩個表現組──通過組與不通過組。所採用的方法包括題目內容法（Angoff, 1971; Ebel, 1972）、對照組法（Koffler, 1980）或是綜合法（Jaeger, 1982）。

九十年代初期，最低能力檢驗的思維，開始有了變化。標準本位的教育改革（standards-based educational reform）帶動了「標準本位評量」（standard-based assessment, U.S. Department of Education, 1996）的興起。評量的目的在探討根據課程所規劃出來的內容標準（content standards），學生在經過學習後，其學業表現是否達到標準？如同 Linn（2000）所說，當代教育改革之核心特色就是創造了標準（standards）。標準指稱兩項主要內容：內容標準（content standards）以及表現標準（performance standards）。內容標準指學生所應理解與所需有的能力，在教育界中通常指稱課程，在我國的課程脈絡中一般認爲是能力指標；表現標準乃指在內容標準的規範下，表現的水準狀況（Hambleton, 2001），亦即學生的表現要多好才算好、好的程度以及品質（Eckhout, Plake, Smith, & Larsen, 2007; Koski & Weis, 2004; Linn, 2000）。二者的關係在於內容標準需先規劃，接著依據內容標準設計標準本位評量，然後根據學生在評量上的表現，設定成就水準，此即爲表現標準（Cizek, 2001）。表現標準一般含有兩個要素：成就水準（achievement levels），以及達到此成就水準的表現品質。例如，美國的教育進

展國家評量系統（NAEP, National Assessment of Educational Progress）採用三個成就水準：基本（basic）、精熟（proficient）、以及精進（advanced），將學生的表現分爲四類：未達基本、基本、精熟、精進（National Center of Education Statistics, 2008）。每類學生的表現品質均有詳細的描述。

過去最低能力檢測時候所盛行的標準設定方法，不容易應用於多個成就水準的情境，書籤標定法便在這一需要下應運而生。此一方法爲 Lewis、Mitzel、與 Green 提出（1996），因其可設定多個切分點以反映多個成就水準，是目前美國最多州所採用的標準設定方法（Perie, 2005; 引自 Karantonis & Sireci, 2006）。本研究之第一項研究目的爲依據能力指標所彰顯的內容標準，實施標準本位評量之後，採用書籤標定法來設定表現標準，以檢核並說明學生的成就。

（二）書籤標定法的項目圖以及達到成就水準的定義

書籤標定法之進行首先需要製作項目圖（item map），重要內涵包括：

- 各題經過項目反應理論（IRT）校正過後之題目難度值；
- 依據難度值將各題由易到難從上到下排序；
- 題目出處；
- 相對應的能力內涵。

項目圖之部分舉例請見第三部分研究方法之表3，其用途乃在協助專家（通常包含學科專家與測驗專家）根據內容標準（即能力指標）所描繪的能力內涵，在由易到難（從上到下）排好序的項目圖中，將書籤插入恰當的題目與題目之間，作爲能力內涵不同成就水準的切分點。所謂達到特定程度的成就水準，定義爲學生必須答對此書籤位置之前所有題目的三分之二題數以上（Lewis et al., 1996）。例如，某個學生要達到「基本」的成就水準，就必須答對「基本」水準之書籤位置之前所有題目2/3以上；換言之，達到「基本」水準的學生，對於難度低於此書籤位置之一系列題目，需有2/3答對率的表現。

三分之二題答對率的訂定，根據 Huynh（1998, 2006）的論述，乃爲「答對題目」之最大訊息量之處。Huynh 提及，二元計分方式（對與錯）的題目之訊息量（亦即題目變異量）爲 $p^*(1-p)$，其中 p 值爲答對的機率。訊息量 $p^*(1-p)$之中，有 p 之比率爲「答對此題」之訊息量，亦即 $p^*(p^*(1-p))$。此訊息量在 p=2/3或是.67時，達到最大。也就是說在答對的可能性爲2/3時，此題之「答對題目」的訊息量達到最大，亦即當答題者有.67的答對機率時，我們對於他答對能力的測量最準確，估計誤差最小。如果每一題都需有.67的答對機率，則就一系列題目而言，則答題者至少需答對成就水準範圍內三分之二的題目數量，才足以說明答題者具有此系列題目所欲測得之成就水準。

（三）書籤標定法的標準設定程序

標準設定即在找出不同成就水準的切分點。切分點意指書籤位置之前的題目難度值所反映出來的能力值。書籤標定法中切分點之設立，需經過三回合以上的討論，每一個回合又必須經過專家個人設定切分點、小組內部相互討論交流、跨小組之間相互討論交流的過程。書籤標定法之一般性步驟（Lewis, Mitzel, Green, & Patz, 1999）如圖1所示：

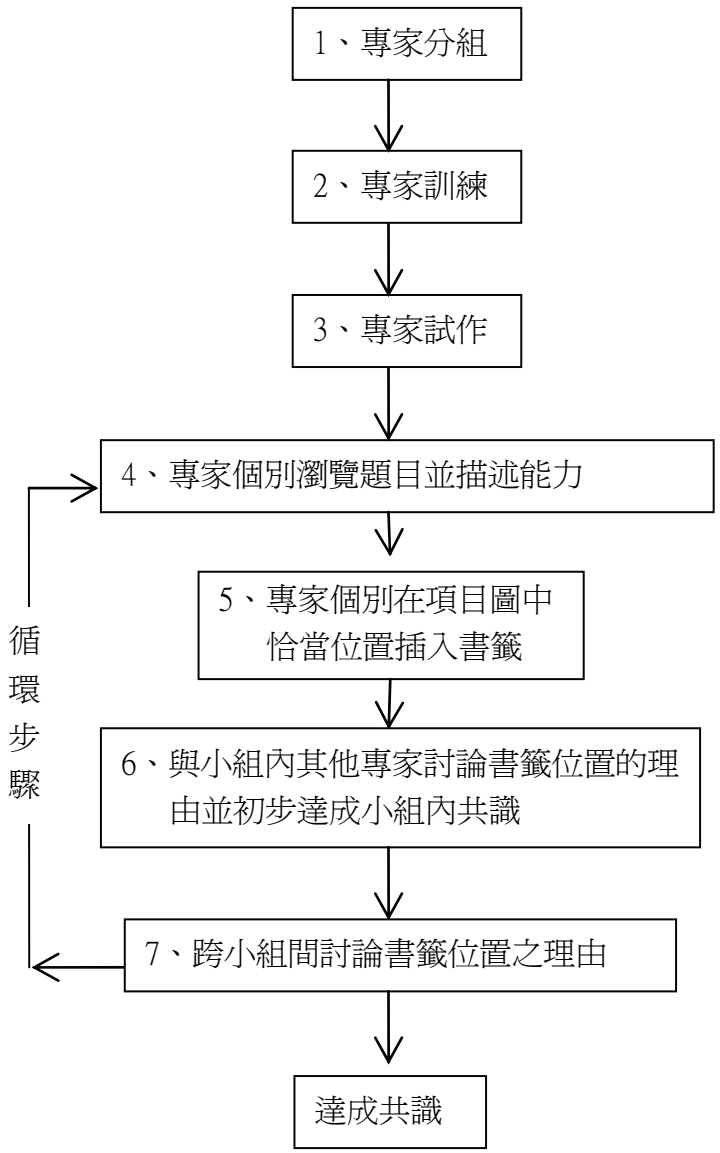

圖1　書籤標準設定法之步驟流程圖

具體說明如下：

1. **專家分組：**專家們通常分成若干小組，每組包含三到五位專家，專家的組成包含學科領域與測驗領域相關專家。

2. **專家訓練：**包含對成員簡介研究目的、測驗性質、能力指標內涵、題目內容、以及書籤標定法的過程及原則。

3. **專家試作：**試驗性地讓專家嘗試插入書籤。專家根據已經由易到難排好序的項目圖中，將書籤插入恰當的題目與題目之間來區隔不同能力組別。

4. **描述能力：**正式開始以後之第一步爲專家閱覽項目圖，並根據自己的專業，提出試題配合能力指標之能力描述。

5. **專家個別設定書籤位置以及切分點：**能力描述出來以後，每一位專家按照自己的判斷，利用項目圖，在恰當的題與題之間，插入書籤，以區分不同能力等級的學生。當學生的能力可以答對書籤位置之前較容易之系列題目數量達三分之二時，便表示該學生具備有此書籤位置界定的成就水準。切分點便是書籤位置之前的題目難度值所反映出來的能力值。

6. **小組內部討論書籤位置：**專家各自定好書籤位置以後，於小組內和其他成員相互討論彼此的理由。討論過後，每一位專家依據討論，修改書籤位置，並統整計算小組內的切分點。

7. **跨小組間討論書籤位置：**以組爲單位進行跨小組討論，說明書籤位置的理由，並且相互說服。討論過後，每一位專家再次修正書籤位置。

8. **循環第4、5、6、7步驟：**跨小組討論過後，循環4、5、6、7步驟。書籤位置經過循環討論後，預計切分點將逐次收斂達到共識。

9. **大會共識或議決：**最後於大會上定案共識的切分點，如果專家之切分點並不一致，通常以中數作爲最終位置，但仍視結果應用之脈絡而定，最後依據切分點所區分出來的試題範圍之內容與難度，標定成就水準，並描述學生能力表現。

上述過程乃爲書籤標定法設定表現標準的流程，完成的任務包含兩項內容：由切分點位置有所反映出來的成就水準之個數，以及各成就水準之表現品質的描述。

（四）書籤標定法的優點

書籤標定法之所以廣爲採用，有其實務上以及學術理論上的優勢，分述如下：

1. 可用於多切分點、多題型情況：書籤標定法之所以在標準本位評量之需要下漸趨熱門，最大的原因乃是書籤標定法的項目圖可混合不同題型的題目，只要計算出題目的難度值，不論二元計分或是多元計分的題目皆可依照難度值混合排在一起。此法亦可適用於多面向能力多切分點的情境（Karantoni & Sireci, 2006; Linn, 2003）。

2. 減少專家的認知負荷：Linn（2003）認爲過去 Angoff 的方法要求專家們必須就不同能力組群學生的可能表現，估計不同能力組群學生對每一題的答對機率，造成專家認知負荷繁重。書籤標定法要求專家根據題目由易到難之順序，估計不同能力表現之切分點所在，整體考量測驗所欲測量的能力，可大量減少專家的認知負荷，專家不必耗盡大部分認知能量在每一個題目上，而得以同時考量整體測驗內容與學生的表現。

3. 結果解釋可融合能力表現：由於專家考量切分點的落點與學生表現之關係，促使參與標準設定的專家對於處在不同能力階段的學生有更清楚的了解。因評量結果的詮釋需綜合學生達到某一水準所具備的知識、技術、能力等，使得結果解釋與學生能力容易相互結合（Linn, 2003）。

4. 可得到較高的專家一致性：與其他方法相較，書籤標定法的一致性較高。Buckendahl、Smith、Impara、與 Plake（2002）以兩組專家針對同一個測驗題本及同一群受試者分別採用 Angoff 方法及書籤標定法進行比較，結果發現書籤標定法的標準誤較小，有較高的評審間同意度。Yin 與 Schulz（2005）亦發現經過二回合討論之後， Angoff 方法中專家的切分點之標準差（10.96）大於書籤標定法中專家切分點之標準差（8.66）。雖然最後兩個方法之切分點差異不大，但是過程中書籤標定法較易促使專家意見一致。

（五）書籤標定法的限制

書籤標定法本身的特性亦造成其應用上的四項限制：

1. 試題難度之排序爭議：書籤標定法之項目圖是由 IRT 相關軟體計算出難度值並依序排列。然而不同領域的專家（例如測驗專家與內容專家），甚至是同一領域的專家審視試題，可能以不同的角度解釋難度排序。有些專家以內容深度解釋題目難度的差異，有些則以認知複雜度來看題目難度差異，這種異質觀點使得專家在插入書籤時增加了書籤所在位置的變異程度。Lewis 等人（1999）認爲

這是使用書籤標定法無法避免的難題，但可藉由專家討論每一試題何以會比前一題難度較高，來提高排序的共識。而 Linn（2003）論及前後緊鄰題目之難度的信賴區間其實有重疊的部分，難易之前後順序確實有討論空間。Skaggs 與 Tessema（2001）研究閱讀測驗題目發現，即便經過討論，仍不能完全避免試題排序的爭議，例如「困難度高的文章中較簡單的閱讀測驗題目」，其排序較「困難度低的文章中較難的閱讀測驗題目」來得前面，造成閱讀題目與閱讀文章分散各處，不容易解釋成就水準的表現狀況。書籤標定法適用於單獨完整的題目，不適於題組或是閱讀測驗形式的題目。

2. 容易忽略試題難度以外的重要資訊：Linn（2003）特別提醒使用者，由於此法僅利用試題難度作爲排序的依據，容易忽略難度以外的其他重要參考訊息。例如，測驗原是測量綜合能力，但由於簡單容易的題目排在較前面的位置，基本的成就水準，只涵蓋了單一能力，（例如，基礎英語能力的內涵變成英語單字能力）。在解釋表現結果時，尤其是低階的能力水準，往往變成單一面向的詮釋。這種以單一能力作爲基本能力的代表，較不符合現實世界中所看到的能力之綜合性。

3. IRT 模式本身的限制：雖然 IRT 有許多好處，但 IRT 的基本假定，卻也限制了書籤標定法的使用。最常見的限制是試題必須符合單維向度（unidimentionality）和局部獨立（local independence）兩大項目反應理論的基本預設，若兩大預設不成立，則標準設定結果之強固性（robustness）將大打折扣（Linn, 2003）。

4. 切分點可能過低：研究發現書籤標定法之切分點低於其他方法之切分點，例如，Yin 與 Schulz（2005）發現書籤標定法低於 Angoff 方法；Green、 Trimble、與 Lewis（2003）發現書籤標定法最低能力組以及最高能力組的切分點，皆低於其他設定方法。造成切分點偏低的理由，Green 等人認爲是因爲專家由簡單的題目開始看起，一旦專家感覺到最低能力組對於題目之答對機率開始低於 .67 時，便立刻提出切分點，造成切分點的負向偏誤（過低）。就此傾向而言，如果相反的把題目由難排到易，專家則會反過來尋找最高能力組學生從哪一題開始答對機率高於 .67，此時，切分點的位置將會偏向較難之題目，造成正向偏誤。Reckase（2006）建議將兩種排序方式得到的兩個切分點間的題目難度相加計算平均以尋求較恰當的切分點。

本研究採取下列步驟以減低書籤標定法的風險：

1. 安排測驗專家與內容專家同一小組，並在正式會議開始之前，以報告發表的方式與小組共同瀏覽全部題目，以促進專家對於試題內容的瞭解。

2. 在設定切分點時，給予專家考慮綜合性能力或是不同能力面向分別進行的彈性（例如分別考量聽能力與讀能力），避免切分點僅考量到單一面向的能力。

3. 研究團隊檢驗試題符合假定之後才採取 IRT 計算各題的題目參數。

4. 關於切分點是否過低之問題，本研究另外採取常態混組模型之心理計量模式的分析，作爲討論有效程度的輻合的證據，以瞭解切分點的適切性。

（六）標準設定有效程度的輻合證據—常態混組模型

輻合證據乃指不同測驗方法測量同一構念時，彼此間的關聯性（American Educational Research Association, American Psychological Association, and National Council on Measurement in Education, 1999; Crocker & Algina,1986）。因輻合證據與效標關聯證據常常採用相關係數，是以兩個證據的性質很容易混淆。效標關聯證據乃在討論某特定工具（例如大學入學指定科目考試）對於一個重要的效標特質（例如，大學時候的學業表現）之有效情況。輻合證據不同於效標關聯證據，乃指兩個不同工具（方法）對於同一特質的測驗結果之關聯，例如同樣測量人際關係，一爲同儕評定法，一爲自我評估報告，二者間的關係便是輻合證據。所輻合的兩個工具同等重要，所得證據可相互支持。至於哪一個工具比較恰當，則視測驗實施與結果使用的情境而言。

本研究以書籤標定法來設定英語聽讀基本能力表現的切分點，至於此切分點如何有效，則需要

效度證據的支持。Hambleton 曾言（2001）表現標準的設定就是一連串人爲判斷歷程，其基本步驟有11項，從專家的選定到成就水準個數的設定，都需要判斷。書籤標定法如上述文獻所整理，亦同樣涉及專家判斷，然與其他方法相較，有其應有便利的優勢。Reckase（2006）認爲人爲判斷歷程之有效性，更需要研究者提供計量效標來檢核之。

在此前提下，本文採用常態混組模型的分類結果來討論書籤標定法有效性之輻合證據，原因乃是常態混組模型與書籤標定法的專家標準設定會議有相同的目的──將學生分成恰當的不同能力組，但二者的理論依據以及方法完全不同。當兩種不同的工具，測量相同構念而得到類似結果時，便是 Campbell 與 Fiske 所提出來的多特質多方法（MTMM, Multi-trait Multi-method）中的輻合效度證據（1959）。

常態混組模型之目的在將一個不同平均值的異質常態混組分佈（heterogeneous normal nixture distribution）區分出數個內部平均值同質的潛在子組別分佈（Basford & McLachlan, 1985; Everitt & Hand, 1981）。常態混組模型的機率密度函數如下：

$$f(x)=\sum_{i=1}^{k}\pi_i\times f_i(x)\quad for\quad 0\le\pi_i\le 1\ ,\ \sum_{i=1}^{k}\pi_i=1 \qquad \text{（式1）}$$

其中 i 表不同潛在子組別，π_i 表各潛在同質子組別佔全體的比率，其機率密度函數爲 $f_i(x)$。透過模式適合度的判斷，研究者由數個競爭模式中選擇適配實徵資料的最佳適配模式。各競爭模式內會有不同數目的潛在子組別，如果以兩個子組別爲例，則其常態混組模型的機率密度函數如下，需估計的參數爲 π，μ_1，σ_1，μ_2，σ_2。

$$f(x)=\pi f_1(x)+(1-\pi)f_2(x)$$
$$=\pi\times\frac{1}{\sigma_1\sqrt{2\pi}}\cdot e^{-(x-\mu_1)^2/2\sigma_1^2}+(1-\pi)\times\frac{1}{\sigma_2\sqrt{2\pi}}\cdot e^{-(x-\mu_2)^2/2\sigma_2^2} \qquad \text{（式2）}$$

當研究者確認適配之模式後，需計算區分潛在子組別的切分點，切分點乃一特定數值 x，使得相鄰潛在子組別的機率密度函數相等，亦即兩個常態分佈的交會處。例如兩個組別時，切分點 x 使得 $f_1(x)=f_2(x)$。我們可對任一觀察值，計算其屬於各組別之事後機率（posterior probability），事後機率最高者便爲此數值應該歸屬的組別。

常態混組模型的典型例子是成年人之身高分佈。今假設有一外星人來到地球，看到地球上的成年人類身高有高有低，他將此身高畫出次數分佈圖，發現此分佈圖是個常態雙峰分佈，於是他猜測地球上的成年人內部應該有兩個潛在子組別。常態混組模型便在計算所觀察的資料是否內含有數個潛在子組別，可用於生物種類的群組探勘，例如特定區域中魚的族群，或是種子之基因型態來自雄株或是母株的可能性（Lindsay, 1995）。本研究則利用此模型來瞭解測驗分數分佈之下是否包含有兩種或三種以上的不同能力組。群集分析（Cluster Analysis）也是一個達成類似目的的分析方式，但差別在於常態混組模型假定平均值同質的潛在子組別爲常態分佈，而群集分析並不假定子組別之分佈狀況；且常態混組模型可利用模型計算特定數值屬於哪一個潛在子組別的事後機率，群集分析則無法提供事後機率。本研究採用 Latent Gold 套裝軟體（Vermunt & Magidson, 2005）分析資料是否具有常態混組模型之性質。

本研究中，受測學生之 IRT 能力值 θ 屬於連續變項，並形成一個分佈，在此分佈下研究者假定具有多個不同平均值的潛在子組別：例如，通過組與不通過組，便是兩個各有自己平均值的潛在同質子組別。正如 Reckase（2006）所呼籲的，人爲判斷的專家標準設定需要計量模式作爲有效性的證據。研究者認爲常態混組模型可解釋上述能力分佈的潛在分組，且以計量模式爲基礎，透過模式與資料的適配程度選擇的分類，適合作爲人爲判斷之標準設定流程的輻合證據。

研究方法

一、研究樣本

爲確保所抽取之樣本具有全國代表性，本研究採兩階段隨機抽樣設計。第一階段爲分層叢集隨機抽樣，第二階段則在所抽取到的樣本學校中，以個人爲單位進行簡單隨機抽樣。各階段詳細抽樣細節請見台灣學生學習成就評量資料庫2005年台灣學生英語學習成就之趨勢調查研究期中報告（陳淑惠等，民94），共抽出10455名學生，經剔除未參與測試以及不適當之受測者（例如資源班學生）354名後，總計10101名有效樣本。

二、英語聽讀基本能力

本研究正式施測的題目如「研究目的與問題」中所言，乃是根據九年一貫課程綱要中之分段能力指標的描述（教育部，民92，民93），就學生學習英語一個學年後（每週2堂課，每堂課40分鐘）的學習狀況而出題，目的在瞭解學生是否達到英語聽讀的基本能力。基本能力之內涵，亦即標準本位評量的內容標準，包括聽懂辨別英文字母的名稱 （letter names）、英語語音（letter sounds）、及發音相近的字詞，同時也能聽辨出英語問句和直述句的語調差異、生活常用語的意義、單一句子的句意、簡易對話的內容以及韻文中關鍵句子的句意。另外在閱讀能力的表現上，學生要能辨識英文印刷體大小寫字母，也能正確了解詞彙以及簡易英文標示的意義，並理解所讀單一句子的句意、韻文中字詞的字義，以及韻文的內容大意。本研究在規劃標準本位評量時，以能力指標爲依據，至於詮釋學生表現有多好的表現標準，包括成就水準的個數、切分點設定、以及表現品質的描述，有賴書籤標定法之標準設定會議討論並共識。

三、平衡不完全區塊（BIB, balanced incomplete block）設計

正式施測所用的題本經過臨床試測並刪除或修改後，由具有代表性之試題所形成，題目之有效性以及構成依據請見陳淑惠等（民95）。每一個正式題本依照平衡不完全區塊（BIB）之設計組成而成。BIB 設計的目的乃緣於題目範圍大，題數多，如讓學生受測所有題目，則耗時也會造成學生疲累，是以依據平均分散原則，將題目分成許多區塊（block），組合不同區塊則可構成不同題本。換言之，題本乃由固定數量的區塊組成，但所用到的區塊各有不同。同一個受測班級學生，會收到不同題本，如此每一個題目都有足夠的受試者接受測試，但每一位受試者又不必接受題庫中全部試題的測驗，負擔較爲輕鬆（Kaplan, 1995）。

本研究英語科資料庫依照題型歸類，共有70個試題，分組成7個區塊（B1-B7），每一個區塊有10個題目，共有3個聽區塊（B1, B2, B3）以及4個讀區塊（B4, B5, B6, B7） 如表1所示。這7個區塊排列組合出正式施測的6個題本，每個施測之題本含4個區塊（聽讀各兩個區塊），共40題試題。聽的題目中，區塊 B2爲共同試題。讀的題目中，則每一個題本與其他題本至少有一個重複之區塊，如題本1、題本2、題本4、與題本5重複區塊 B5；題本2、題本3、題本5、與題本6重複區塊 B6，以利定錨分析之用。每個學生僅會被測到六個題本中之一本（亦即40題），各題本所含區塊之排列見表1：

表1　施測題本不完全平衡區塊設計

題本號	聽題之區塊		讀題之區塊	
題本 1	B1	B2	B4	B5
題本 2	B1	B2	B5	B6
題本 3	B1	B2	B6	B7
題本 4	B2	B3	B4	B5
題本 5	B2	B3	B5	B6
題本 6	B2	B3	B6	B7

本研究根據題目反應理論之模式適配狀況，選定三參數（包含猜測機率）的估計模式。由於透過試題特徵曲線函數可以計算學生在該題的答對率，故研究者藉由 BILOG-MG 統計軟體估計出每一位學生的能力值之後，再代入70個試題的試題特徵曲線（ICC）估計該生在該題的答對率。本研究樣本在70個試題的平均估計答對率是0.85。本研究所使用之試題難度屬於「中等偏易」。

四、標準設定程序

依據教育與心理測驗歷程標準（AERA, APA, & NCME, 1999, p. 59）的建議，解釋及使用切分點時，關於評斷該切分點的方法、擔任評鑑的專家、專家們訓練的過程以及討論的步驟、試題的表現等資料，均應詳細提供，這就是所謂程序證據的意涵，亦爲本研究標準設定是否有效的重要指標之一。以下，研究者將比較文獻提及的書籤標定法程序以及實際運作的書籤標定法程序，加以對照及討論。

（一）專家組成

英語科資料庫標準設定會議之組成有13位專家，包括3資深國小教師、7位英語領域研究者、以及3位測驗領域研究者，另外並有3位研究助理。全部專家分成3個小組，每個小組包含一至二位資深國小教師、兩到三位英語科領域教授、以及一位測驗領域教授，另外每一組有一位研究助理協助記錄。

（二）會議時程

標準設定會議的時程，依據文獻中書籤標定法的步驟所制定。會議時程具體呈現如表2。

表2　標準設定會議時程

時間	活動主題	活動內容
1:30-2:10	標準設定說明與釋疑	澄清設定之目的並概覽試題
2:15-3:00	第一回合小組討論	1.記錄個人切分點及相關定義
		2.小組討論與調整切分點
		3.小組共識並記錄
		4.推派發言人
3:10-3:30	第一回合大會討論	5.小組結果分享（切分點範圍）
		6.大會提供並解釋題目特質
		7.大會提供並解釋學生表現
		8.大會調整切分點並記錄
3:40-4:00	第二回合小組討論	重複第 1 到第 4 步驟
4:30-5:00	第二回合大會討論	重複第 5 與第 8 步驟
5:00-5:20	第三回合小組討論	重複第 1 到第 4 步驟
5:20-5:40	第三回合大會討論	重複第 5 與第 8 步驟
5:40-5:45	大會共識切分點之位置	
5:45-6:00	結論	

（三）訓練材料

首先，大會在標準設定說明與釋疑時候，給予專家項目圖。所有題目依據難度之高低，由易到難從上到下排列。部分題目所排成之項目圖舉例如表3：

表3　英語聽讀能力標準設定之項目圖（item map）舉例

題號	內容	認知	主題	施測人數	答對人數	通過率	二系列相關	鑑別度	難度	猜測度
E402	音素/字母	辨識	辨識印刷體大小寫字母	3386	3308	0.98	0.82	2.07	-2.76	0.24
E702	音素/字母	辨識	辨識印刷體大小寫字母	3362	3273	0.97	0.81	2.01	-2.64	0.24
E502	音素/字母	辨識	辨識印刷體大小寫字母	6777	6597	0.97	0.89	2.38	-2.59	0.13
E602	音素/字母	辨識	辨識印刷體大小寫字母	6753	6570	0.97	0.89	2.27	-2.56	0.17
E401	音素/字母	辨識	辨識印刷體大小寫字母	3386	3287	0.97	0.88	2.16	-2.54	0.21
E501	音素/字母	辨識	辨識印刷體大小寫字母	6777	6577	0.97	0.90	2.36	-2.50	0.14

本次訓練材料之重點包括六項內容：TASA 資料庫之目的、英語聽讀基本能力指標（教育部，民93）、雙向細目表（見陳淑惠等，民94）之瀏覽、BIB 試題之設計、參數估計結果的解讀、書籤標定法的精神與流程、以及項目圖（如表3舉例，包括題目內容）。依據書籤標定法之步驟（見文獻探討），從步驟1到步驟4，在大會充分提供能力指標內容以及雙向細目表資料下，皆進行得相當順利。但在步驟5處，由於插入書籤的思維模式與英語科內容專家過往所熟悉的內容關聯證據之專家評定方式有所不同，許多專家剛開始接觸書籤標定法時顯得不易適應，因此會議說明及釋疑部分花去較長時間。主要說明的議題爲：

1. 插入書籤設立能力切分點考慮的是學生能力表現的分界點而不是試題內容之歸類。

2. 書籤位置所顯現的意義乃是學生如能答對書籤位置前面的題目數量達2/3（題目由易到難排序），則表示該學生具有此系列題目所測量的能力。

3. 專家誤以爲2/3是指「全體學生」對於此測量結果需有2/3的通過比率。大會再度與專家說明，2/3答對機率乃指一個學生如具有基本水準，則必須能夠答對基本水準題目中三分之二以上的題目。例如，基本水準的題目共30題，則2/3答對率指的是學生能夠答對這30題中的20個以上的題目。專家需注意書籤位置乃在反映該位置之前的題目內容符合該位置所要表達的成就水準。

（四）三回合循環討論

在第5步驟時，大會解決了專家的疑惑之後，便順利進入第6、7步驟以及第8步驟的循環。標準設定會議一共有三回合循環討論，每一位專家共有紀錄單一式三份。專家紀錄單中包含組別、專家姓名、各個切分點在項目圖上之難度排序的位置、設立該切分點之理由、以及能力特質之描述，具體呈現如表4。由於本測驗在能力指標之目的下，屬中等偏易之題目（平均估計答對率爲.85），對於能力高之學生的鑑別度不足，故不考慮四個能力組別的存在。研究者在專家的記錄單上提供兩個切分點之空格，目的在給予專家討論組數之彈性，專家可視學生表現狀況，將學生能力分成兩組（一個切分點），或是三組（兩個切分點）。

表4　標準設定小組或專家個人紀錄單

第＿＿＿回合　□小組用　□專家個人用（專家姓名：＿＿＿＿＿＿）

較低能力到中間能力切分點位置：＿＿＿＿＿＿

理由：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

大會提供資料後請填上學生表現累計百分比：＿＿＿ 題目難度界於：＿＿＿

中間能力到較高能力切分點位置：

理由：＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿＿

大會提供資料後請填上學生表現累計百分比：＿＿＿ 題目難度界於：＿＿＿

三類學生能力標籤與特質描述：

	低能力組	中間能力組	高能力組
標籤			
特質描述			

文獻建議在正式開始設定切分點以前，需要一次「試作」，但由於本次書籤標定法之資料爲研究團隊首次蒐集得來，手邊並沒有恰當的真實資料與題目讓專家實際試作，本研究乃將試作包含在最開始「標準設定說明及釋疑」，並以舉例的方式進行。

本研究書籤標定法之實施程序，大致上符合文獻探討中所建議的書籤標定法之標準設定程序。至於專家三個回合的判定過程、依據、與共識，以及輻合證據所提供的有效性狀況，請見以下結果。

研究結果

一、標準設定判斷歷程與結果

（一）標準設定過程

標準設定會議經過三個小組三回合討論，13位專家根據基本能力內涵以及學生表現狀況，設立表現標準之切分點以反映成就水準，判斷過程如下（見表5）：

1. 在第一回合討論結束時，發生兩個現象：

（1）11位專家（專家 C 到 M）根據題目偏易的性質以及基本能力內涵，認爲學生表現分爲兩組便可，無法再區分出精進組，但聽與讀能力應分開，各設一個切分點；

（2）另外2位專家（專家 A 與 B）則傾向將聽讀合併，設立兩個切分點。

2. 第二回合時，三組專家的看法沒有變動，和第一回合結果相同。

3. 第三回合時，專家 A 與 B 改變想法，認同設一個切分點。

三個回合切分點的數目以及題目難度值見表5。

表5　書籤標定法十三位專家三回合切分點之位置

		第一回合			第二回合			第三回合	
		切分點 1		切分點 2	切分點 1		切分點 2	切分點 1	
		聽	讀		聽	讀		聽	讀
第	A 專家	-1.45		-0.14	-1.45		-0.14	-0.49	
一	B 專家	-1.45		-0.53	-1.45		-0.53	-0.59	
組	C 專家	-0.55	-0.22	不需要	-0.55	-0.22	不需要	-0.55	-0.22
	D 專家	-0.59	-0.14	不需要	-0.59	-0.14	不需要	-0.59	-0.14
第	E 專家	-0.59	-0.14	不需要	-0.59	-0.14	不需要	-0.59	-0.14
二	F 專家	-0.59	-0.14	不需要	-0.59	-0.14	不需要	-0.59	-0.14
組	G 專家	-0.59	-0.14	不需要	-0.59	-0.14	不需要	-0.59	-0.14
	H 專家	-0.59	-0.14	不需要	-0.59	-0.14	不需要	-0.59	-0.14
	I 專家	-0.49	-0.55	不需要	-0.49	-0.55	不需要	-0.49	-0.55
第	J 專家	-0.49	-0.55	不需要	-0.49	-0.55	不需要	-0.49	-0.55
三	K 專家	-0.49	-0.55	不需要	-0.49	-0.55	不需要	-0.49	-0.55
組	L 專家	-0.49	-0.55	不需要	-0.49	-0.55	不需要	-0.49	-0.55
	M 專家	-0.49	-0.55	不需要	-0.49	-0.55	不需要	-0.49	-0.55

（二）聽讀合設或分設之判斷

經過三個回合討論後，13位專家有共識的部分爲：設1個切分點。至於聽讀合併或是分開考量，有2位專家（A 專家與 B 專家）贊成聽讀合併，11位認爲應該聽讀分開計算。值得注意的是，第三組專家雖將聽讀能力分別切在聽試題代號 E109（難度值-0.49）和讀試題代號 E405（難度值-0.55）兩個題目上，但仔細考量切分點所在位置的題目難度，其實這兩個題目的難度極爲接近。經提出此現象後，第三組專家認爲聽讀可以合併切分點。

此時有共識的部分爲全部專家認爲設一個切分點，其中7位專家（A、B、I、J、K、L、M）贊成聽讀合設，其餘6位專家認爲應該聽讀分設。後又經過討論，最後共識聽讀合併設一個切分點，理由如下：

1. 本次評量試題偏易，專家認爲設立一個切分點表示「通過」與「不通過」便已足夠，實不容易再區分出英語基本能力之精進組。

2. 如果依照第二組專家的建議，聽讀分開設置切分點，則學生表現會出現四種組合：聽讀都過，聽讀都不過，以及人數很少的「聽過、讀不過」，因爲聽與讀之切分點的試題難度值差異不大（見表5，聽切分點難度值-0.59；讀切分點難度值-0.14），還有人數更少的「聽不過、讀過」，在解釋上益增困擾。另外，以美國 NAEP 的經驗（National Center for Education Statistics , 2008）爲例 ，學科成就測驗包含不同內容，但在分數解釋上，亦統整爲一個綜合性能力。

此時專家們共識聽讀合設一個切分點，候選切分點的試題難度值在第三回合時得到專家認可的次數分配分別爲：-0.22一次、-0.14五次、-0.49六次，-0.55六次，-0.59六次（見表5），顯然集中在-0.5與-0.6之間，這三題試題代號爲 E109、E405、E201。項目圖上的難度值乃爲學生答對該題機率.5時的試題難度值，依據本文文獻所述（Huyhn, 1998, 2006），切分點通常選取通過率.67（即2/3）時的能力值。因此，將該三題之 IRT 試題參數代回 IRT 之三參數公式，求取這三題在 p=.67時之學生能力值。IRT 三參數之公式如式3所示：（Hambleton & Swaminathan，1985）

$$P_i(\theta) = c_i + (1-c_i)\frac{\exp\left[a_i(\theta - b_i)\right]}{1+\exp\left[a_i(\theta - b_i)\right]} \qquad （式3）$$

三題之參數值如表6：

表6　切分點候選試題之 IRT 參數與答對機率.67時的能力值

題目	鑑別度 a	難度 b	猜測度 c	答對機率在.67 時的能力值 θ
E109	4.13	-0.49	.43	-.57
E405	3.27	-0.55	.36	-.56
E201	1.99	-0.59	.19	-.40

（三）三個候選試題之能力切分點判斷與能力描述

在答題者答對該題機率 *p*=.67時，三個候選試題 E109、E405、E201所反映之能力值分別爲-.57，-.56，-.40。由於-.57與-.56相當接近，代表較多數的結果，且考量標準本位評量爲我國第一次施行，專家們皆認爲取其中較低的能力值便可，最後大會共識採用-0.57作爲台灣學生英語文學習成就評量學生通過與否之切分點。亦即，學生能力值大於等於-0.57爲「通過」，有72.9%的學生，其英語能力具有國民小學英語文學習領域能力指標之水準（教育部，民92，民93），在計量意涵上則是通過的學生至少有2/3的機率能夠答對在難度上比這切分點更容易的題目。

通過的學生具有英語基本能力，其表現品質包含以下兩個要素：

1. 就聽而言，學生大致上能夠聽懂辨別英文字母的名稱（letter names）、英語語音（letter sounds）、及發音相近的字詞之外，也能聽辨出英語問句和直述句的語調差異、生活常用語的意義、單一句子的句意、簡易對話的內容以及韻文中關鍵句子的句意。

2. 就讀而言，學生大體能辨識英文印刷體大小寫字母，也能正確了解詞彙以及簡易英文標示的意義，並能理解所讀單一句子的句意、韻文中字詞的字義，以及韻文的內容大意。

能力值小於-0.57者爲「不通過」，表示經過一個學年的學習後，學習表現在程度上還不能有效掌握上述基本能力。至於是哪些部份未能掌握，則需視個別學生答題狀況而定。

二、常態混組模型之輻合證據探討

本研究使用常態混組模型做爲標準設定輻合證據之依據。首先研究者將學生的作答反應透過 BILOG-MG 軟體，採用三參數項目反應模式配合邊際最大概似估計法（MMLE）得到10101個學生的能力估計值，以此作爲模式探究的樣本。

由於標準設定專家會議進行之時，專家們曾對設定一個切分點（將學生分爲兩種能力狀況）或設定兩個切分點（將學生分爲三種能力狀況）進行過討論。本研究配合專家考慮過程、以及九年一貫課程基本能力之目的（在瞭解學生基本能力之狀況而不在區分成就高下），研究者逐一假定學生樣本中有一個、兩個、或三個潛在子組別模式時各模式的適配狀況。結果如下：

表7　不同潛在子組別數目之模式適配度

	結構狀況	Likelihood	AIC
模式一	內含一個潛在組之模式	-14287.65	28579.29
模式二	內含二個潛在組之模式	-13589.43	27186.85
模式三	內含三個潛在組之模式	-13351.13	26714.25

（一）模式選擇——模式三最佳

由上表可知，模式三的適配度比模式一與模式二來得好，AIC 值最低（26714.25＜27186.85＜28579.29），顯示3個潛在子組別的常態混組模型比另外兩個模式更能解釋資料的狀況。AIC 值的計算方式（Akaike, 1987）如下：

$$AIC = \chi^2_{\text{模式}} - 2 \cdot df_{\text{模式}} \qquad \text{（式4）}$$

其中$\chi^2_{模式}$表示此模式適配資料之卡方值，$df_{模式}$爲模式的自由度。AIC 乃在比較哪一個模式具有較小的卡方值以及較大的自由度，AIC 值越小表示相對而言其模式適配與模式精簡（model parsimony）狀況較佳，爲模式選擇的重要參考指標。三種模式的各潛在子組別之平均數、佔母群分佈之比率見表8，模式二將能力分爲極弱組（約百分等級3以下）以及非極弱組，無法描述學生的英語聽讀基本能力之表現狀況。模式三的3組能力分組狀況以及 AIC 值顯示其爲最佳模式。

表8　常態混組模型三種模式下各模式潛在子組別所佔比率與能力平均值

	模式一： 存在 1 個潛在子組別	模式二： 存在 2 個潛在子組別		模式三： 存在 3 個潛在子組別		
		組 C	組 B	組 C	組 B	組 A
占全體比率	1	0.03	0.97	0.02	0.33	0.65
平均值	-0.08	-3.19	0.83	-3.57	-0.86	0.44
標準差	1.00	0.83	0.83	0.59	0.59	0.59

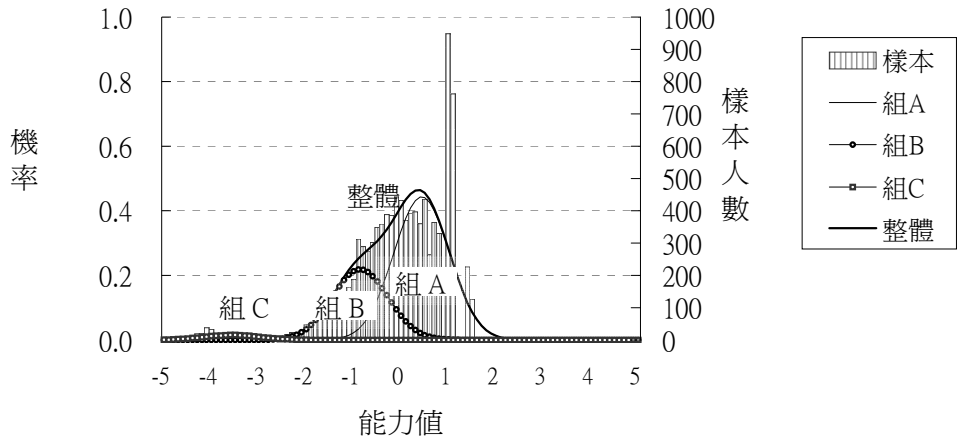

圖2　「模式三」之三潛在組的常態混組模型分佈狀況

圖2之分佈圖顯示模式三中3個潛在子組別 A，B，與 C 之能力值的分佈，其中有兩個能力值之切分點：一爲組 A 與組 B 常態分佈交會點 θ1，另一爲組 B 與組 C 常態分佈的交會點 θ2。根據常態分佈機率密度函數如式5：

$$f(x)=\frac{1}{\sigma\sqrt{2\pi}}e^{-(x-\mu)^2/2\sigma^2} \qquad (式5)$$

以及表8所示的組 A、B、與 C 的平均值與標準差，可知當 θ1＝-0.40時，組 A 與組 B 的機率密度函數相等（兩組常態分佈交會處）；當 θ2=-2.55時，組 B 與組 C 的機率密度函數相等，θ1與 θ2便是不同能力狀況的切分點。全體學生中有6805名學生（佔67.4%）之能力值在 θ1= -0.40之上。另外有3076名學生（佔30.5%）能力位於 θ1到 θ2之間，屬於基本能力未通過組。最後，還有220名學生在 θ2=-2.55切分點之下，佔全體學生2.2%，屬於基本能力極弱組。由於本研究之動機在瞭解一般學生學習英語一個學年後（每週兩堂課，每堂40分鐘）英語聽讀基本能力的狀況。在此動機下，整體測驗中間偏易，就描述學生的表現而言， -0.40比-2.55更適合作爲能力通過之切分點，通過的學生佔全體67.4%，不通過的學生佔全體32.7%，能力範圍包含 θ2=-2.55之能力極弱切分點。關於 θ1與 θ2兩個切分點與書籤標定法的切分點之分類一致性狀況，請見以下 Kappa 分析。

（二）書籤標定法之專家結果與常態混組模型的 Kappa 一致性檢定

書籤標定法之通過組與不通過組之區分爲 θ= -0.57，常態混組模型得到兩個切分點，θ1= -0.40，θ2= -2.55。研究者列聯書籤標定法與 θ1的分類一致性結果如表9，以及與 θ2的分類一致性結果如表10，並計算 Cohen's Kappa 指標，以決定書籤標定法標準會議結果與常態混組模型之哪一切分點的決策一致性狀況較佳。Kappa 之計算方法如下（Swaminathan, Hambleton, & Algina, 1974; 余民寧，民91；張郁雯，民93）：

$$\kappa = \frac{\sum O_{ii} - \sum E_{ii}}{n - \sum E_{ii}} \qquad （式6）$$

其中 O*ii* 爲觀察到的一致性總次數， E*ii* 爲兩種方式因機運而產生的期望一致性。Kappa 的計算，乃從兩種方法的一致性總次數中扣除自然機運產生的期望次數，使得該係數已排除機運造成的一致性膨脹現象。根據式6，表9之 Kappa 值 .87乃得自如下之計算過程： .87=(2733+6805-E)/(10101-E)，E 爲5855.47＝10101*（.271*.326+.729*.674）。表10之 Kappa 值 .11亦採用相同計算流程。

表9 書籤標定法切分點（*θ*= -0.57）與常態混組模型組 A 與組 B 切分點（*θ*1= -0.40）之決策列聯表

		常態混組模型結果（θ1= -0.40）		
		不通過	通過	總和
書籤標定法切分點（*θ*= -0.57）	不通過	2733	0	2733（27.1%）
	通過	563	6805	7368（72.9%）
	總和	3296（32.6%）	6805（67.4%）	10101（100%）

Kappa= .87 （*p =.000*）

表10 書籤標定法切分點（ *θ*= -0.57）與常態混組模型組 B 與組 C 切分點（ *θ*2=-2.55）之決策列聯表

		常態混組模型結果（θ2= -2.55）		
		不通過	通過	總和
書籤標定法切分點（*θ*= -0.57）	不通過	220	2513	2733（27.1%）
	通過	0	7368	7368（72.9%）
	總和	220（2.2%）	9881（97.8%）	10101（100%）

Kappa=0.11 （*p=.000*）

比較兩個表格之 Kappa，可知切分點 θ1（組 A 組 B 交會處）與專家判定結果之 Kappa 達到 .87，比 θ2 （組 B 與組 C 交會處）與專家判定結果的一致性（Kappa= .11）來得高，根據 Sim and Wright（2005）以及 Shrout（1988），此數值表示一致性結果非常好。常態混組模型結果顯示能力值低於-0.40者爲能力不通過，此值亦包含了能力非常弱的另一切分點-2.55。至於影響 Kappa 值高低的盛行率（prevalence）以及偏差率（bias）將於討論中進一步論述。

結論與討論

TASA-EN 乃爲標準本位評量，依據內容標準（即能力指標）設計評量進行施測，透過書籤標定法訂定表現標準，並據此標準檢視學生成就狀況，且進一步以混組模型之計量證據確認所設標準之有效性。本研究 TASA-EN 經過書籤標定法的專家判斷及共識後，表現標準有一個成就水準，將學生

分成兩類─通過與否，切分點爲能力值-0.57。研究者利用常態混組模型之計量模式切分點與之相比較，得到很高的 Kappa 分類一致性 .87。兩種工具之目的相同，但是判斷歷程不同：書籤標定法依靠專家對於內容標準、表現標準、以及成就水準的熟悉而共識出標準；常態混組模型選擇最佳模式並計算適合的切分點。不同方法但測量相同特質的工具得到輻合的結果，表示書籤標定法之專家判斷得到計量模式的有效支持。關於此結果與研究歷程，有以下議題值得討論：

一、既然書籤標定法得到常態混組模型輻合證據之支持，為何不直接採用較經濟的計量模式？

在標準設定的目的下，兩種方式都在對樣本進行判斷以推論母群狀況，兩種方式都無法斷言自己的判斷結果較接近真實。標準設定之所以需要專家，乃因專家同時考量試題難度、性質、受試者能力、社會情境、與測驗結果之用途，故其所設定的結果，加上了脈絡因素，可符應實用現場的需要。例如美國 NAEP 之成就水準區分出4個組別，但計量模式不一定能夠得出4組別之模式，而4個組別的分法卻是經過考量後對於社會大眾而言較適切的測驗結果解釋方式。本研究的專家最後綜合聽與讀之考量，這也是單純計量模式無法思考的決策。至於專家設定的標準（切分點）有效性如何，便需要不同來源的證據支持，此亦爲本研究之意圖。

二、基本能力的表現標準，為何未在測驗編製前擬定好？

表現標準不同於內容標準，並非測驗編製前便先擬好。我國標準本位評量開始實施不久，本研究之專家對於內容標準，亦即能力指標與基本能力內涵，皆已有共識，但關於表現標準之成就水準數目與切分點訂定，則經過一番討論。Linn 曾語重心長提及（2000）在績效責任之壓力下，各州之標準本位評量所設定的表現標準漸漸趨向高標準（high standards），然而此趨勢將導致學生的教育經驗窄化、越來越多的失敗者、甚至於限制了有才華學生的發展。相對而言，就需要補救教學的學生而言，找出哪些學生還需要進一步的補救，才是更急切之道，這便是 Linn 所疾呼的一般標準（common standards）的重要性。而此呼籲亦顯現出奠基於相同內容標準的標準本位評量，其表現標準則常因社會期許而有變動，並不是在測驗編製前已先擬定。或有表現標準已先擬定，如美國之 NAEP，那亦是發展了38年後的結果（Vinovskis, 1998)。TASA-EN 的編製從內容標準出發，到標準本位評量施測後的表現標準，展現出一致性的精神──目的不在於高標準爲何，而在於一般標準狀況下，哪些學生通過基本能力，而哪些學生不通過，尚未具有基本英語聽讀能力。至於這樣的標準本位評量的難度，以及如此表現水準的設定，是否代表未來九年一貫課程綱要中之英語文領域學習成就檢核便以此爲依歸？這是研究者無法確認的議題，如同 Cizek 所言，「標準設定也許算是心理計量方法，但比起其他方法，標準設定無疑是其中最藝術、政治、以及文化的混合。」（2001, p. 5）

三、三組專家中有兩組專家組內意見完全一致，為何如此？

關於專家意見的趨同性，本研究第二組與第三組專家在第一回合「個別專家設定切分點」的階段中，組內專家便已趨向一致，並且在後續的小組討論以及各回合循環討論中，即便有變動，亦是組內專家之意見一起變動。在設定會議開始之前，專家們都瞭解可以有自己的看法，並知悉之後將會透過討論相互協調。至於本研究發生的趨同性高之現象，經過會後與專家訪談，有以下數點可能原因：

（一）在任務結果有重大後果影響的情況下，專家們傾向不突顯個別性，以保持群體中的共識。

（二）專家們第一次接觸書籤標定法，因爲不熟悉，故而傾向將個人看法揉合入組內共同意見之中。

（三）試題較簡單，專家們的見解容易趨向同質，很快就共識出具代表性的切分點。

（四）座位安排方式影響意見的趨同性，當專家個別設定標準時小組聚集同坐一桌，使得個別所設的切分點因爲相互討論而相同。

欲解決上述狀況，未來專家會議首先宜注意座位安排，在第5步驟專家個人插入書籤時候，給予獨立空間以儘量保留專家之個人專業判斷。其次，專家會議中，充分提供專家試作的機會，以讓專家熟悉程序並對程序有信心。同時大會亦需提供專家關於評量結果將如何被運用的說明，以避免專家以爲此結果相當嚴重產生心理壓力或是以爲此結果無關緊要而過於輕忽的誤解。另外，關於專家意見趨向一致的現象，Yin 與 Schulz（2005）亦有類似發現，書籤標定法之專家所設的切分點之標準差在第二回合時候便進一步縮小，輻合程度加大，書籤標定法的專家共識狀況勝過 Angoff 方法。

四、書籤標定法之切分點低於常態混組模型之切分點，需要特別注意嗎？

本研究結果顯示書籤標定法的切分點略低於常態混組模型所得，與過去的文獻謂書籤標定法較容易得到負向偏誤的切分點相符（Green et al., 2003; Yin & Schulz, 2005）。爲了避免此現象，未來使用書籤標定法時，可以提供專家兩種項目圖：一爲「由易而難」的排列方式，另一爲「由難而易」的排列方式，或者透過兩批專家分別利用不一樣的項目圖設定切分點，以資相互參考。然而如此做的缺點便是討論耗時，同時亦耗損經費與人力。因此，研究者首先需判斷測驗結果的嚴重性，在切分點適切性以及資源耗費之間尋找平衡。

五、關於 Kappa 的一致性高低，如何考慮相關因素的影響？

影響 Kappa 一致性的高低有三個因素：盛行率（Prevalence）、偏差率（Bias）、以及不獨立性（Nonindependent Ratings）（Sim & Wright, 2005）。盛行率乃指分類一致的情況中，正向一致性與負向一致性（positive and negative classification）差距的絕對值相對於整體的比率。本研究之盛行率依據表9之數據計算後爲 .4=（6805-2733）/10101。盛行率愈高表示 Kappa 受機運影響越大，如研究者認爲盛行率過高，研究者可進行盛行率之校正，目的在估計更爲務實的 Kappa 值。校正的方法爲將一致性次數平均分配給正向一致性與負向一致性二分類結果中（Sim & Wright, 2005）。以本研究爲例，乃是將原先6805與2733平均之值4769代入式6中，得到 .89=(4769+4769-E)/(10101-E)，E 爲5034.66＝10101*（.472*.528+.528*.472）。由於本研究的盛行率不算太高，且校正之後，原數值從 .87調整爲 .89，亦影響不大，表示本研究 Kappa 值在校正之前亦已恰當排除機運影響，值得信賴。

偏差率指分類不一致情況下，正負向兩種不一致次數差距絕對值對於整體次數的比率，偏差率越高者會導致 Kappa 一致性降低。本研究之偏差率計算結果爲：（563-0）/10101＝.06，幾乎沒有偏差狀況發生，無須進行偏差率校正。

關於「不獨立因素」，乃指兩種分類方式乃來自類似的模式或是類似的思考流程，造成邏輯上循環論證之不當。分類方法是否獨立乃根據學理判斷，本研究兩種分類方式，一爲計量模式一爲專家判斷，二者進行過程彼此獨立，可排除「不獨立性因素」對 Kappa 值的負面影響。本研究書籤標定法的分類結果與常態混組模型之分類結果達到 kappa .87之一致性，綜合上述三個因素之考量，此數值可以信賴，且顯示專家判斷的書籤標定法，得到計量模式之輻合證據有效性的支持。

六、以題目難度之排序為唯一的判斷，如何克服過於狹隘的考量？

項目圖以題目之難度爲唯一排序依據，會產生基礎能力切分點在詮釋上的困難（Linn, 2003）。本研究在設定會議中，有專家提出質疑，謂項目圖中排序前面的試題幾乎都是單字類，如果切分點置

於此處，是否代表只要會單字就達到英語科基本能力？在英語學習歷程中，聽與讀或是單字與句子之理解是同時並進的，單字能力本身不足以說明綜合性的英語聽讀基本能力。為解決此問題，本研究13位專家中有11位專家將聽與讀分開，分別設定基本能力切分點，一直到最後階段，才綜合聽與讀之不同切分點。據此，最後所得到的切分點，便可同時包含英語聽與讀的題目。建議未來研究者進一步比較兩種方法：一則全部過程以綜合性能力設立切分點，一則以「先分面向後綜合」的方式設立切分點，不知何者較為有效？

書籤標定法在標準設定上，廣受歡迎，然有效化歷程之計量證據在文獻上並不充分。本研究首次提出常態混組模型對於書籤標定法所設切分點的輻合證據，並系統性整理標準設定的歷程與問題。一則希望在實務上，分享所經歷的問題與解決的經驗。另外，在學理上，雖然此刻本研究必須就目前的證據，討論書籤標定法所設切分點具有常態混組模型之輻合證據，然有效化是一永遠不斷的探索與辯證，如同本研究所提出來的六項議題，未來如何精進標準設定之方法，仍有待繼續探究。

參 考 文 獻

余民寧（民91）：教育測驗與評量─成就測驗與教學評量。台北：心理。

張郁雯（民93）：信度。載於王文中、呂金燮、吳毓瑩、張郁雯、張淑慧合著：教育測驗與評量（95-128頁）。台北：五南。

教育部（民89）：國民小學九年一貫課程暫行綱要。台北：作者。

教育部（民92）：國民小學九年一貫課程綱要。台北：作者。

教育部（民93）：英語文學習領域能力指標解讀與示例手冊。台北：作者。

陳淑惠、吳毓瑩、何東憲、張郁雯、陳錦芬（民94）：台灣學生學習成就評量資料庫2005年台灣學生英語學習成就之趨勢調查研究期中報告。台北縣：國立教育研究院籌備處。

陳淑惠、吳毓瑩、張郁雯、何東憲（民95）：台灣學生學習成就評量資料庫2005年台灣學生英語學習成就之趨勢調查研究技術報告。台北縣：國家教育研究院籌備處。

Akaike, H. (1987). Factor analysis and AIC. *Psychometrika, 52*, 317-332.

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

Angoff, W. H. (1971). Scales, norms, and equivalent scores. In R. L. Thorndike (Ed.), *Educational measurement* (2nd ed., pp.508-600). Washington, DC: American Council on Education.

Basford, K. E., & McLachlan, G. J. (1985). Likelihood estimation with normal mixture models. *Applied Statistics, 34*, 282-289.

Buckendahl, C. W., Smith, R. W., Impara, J. C., & Plake, B. S. (2002). A comparison of Angoff and bookmark standard setting methods. *Journal of Educational Measurement, 39*(3), 253-263.

Campbell, D. T., & Fiske, D. W. (1959). Convergent and discriminant validation by the multitrait-multimethod matrix. *Psychological Bulletin, 56,* 81-105.

Cizek, G. J. (2001). Conjectures on the rise and call of standard setting: An introduction to context and practice. In G. J. Cizek (Ed.). *Setting performance standards: Conxepts, methods, and perspectives* (pp. 3-18). Mahwah, NJ: Lawrence Erlbaum Associates.

Crocker, L., & Algina, J. (1986). *Introduction to classical and modern test theory.* NY: Holt, Rinehart and Winston.

Ebel, R. L. (1972). *Essentials of educational measurement*. NJ: Prentice-Hall.

Eckhout, T. J., Plake, B. S., Smith, D. L., & Larsen, A. (2007) Aligning a state's alternative standards to regular core content standards in reading and mathematics: A case study. *Applied Measurement in Education, 20*(1), 79–100.

Everitt, B. S., & Hand, D. J. (1981). *Finite mixture distributions*. London: Chapman and Hall.

Flanagan, J. C. (1951). Units, scores, and norms. In E. F. Lindquist (Ed.), *Educational measurement* (1st ed., pp. 695-763). Washing, DC: American Council on Education.

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.

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

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principle and application*. Massachusetts: Kluwer Academic Publishers.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion referenced interpretation. *Journal of Educational and Behavioral Statistics, 23,* 35-56.

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.

Jaeger, R. M. (1982). An iterative structured judgment process for establishing standards on competency tests: Theory and application. *Educational Evaluation and Policy Analysis, 4*, 461-476.

Jaeger, R. M. (1989). Certification of student competence. In R. L. Linn. (Ed.), *Educational measurement* (3rd ed., pp. 485-514). NY: American Council on Education/Macmillan.

Kaplan, D. (1995). The impact of BIB *s*piraling-induced missing data patterns on goodness-of-fit tests in factor analysis. *Journal of Educational and Behavioral Statistics, 20* (1), 69-82.

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

Koffler, S. L. (1980). A comparison of approaches for setting proficiency standards. *Journal of Educational Measurement, 17,* 167-178.

Koski, W. S., & Weis, H. A. (2004). What educational resources do students need to meet California's educational content standards? A textual analysis of California's educational content standards and their Implications for basic educational conditions and resources. *Teachers College Record, 106*(10), 1907–1935.

Lewis, D. M., Mitzel, H. C., & Green, D. R.(1996). *Standard setting: A bookmark approach.* Symposium presented at the Council of Chief State School Officers National Conference on Large Scale Assessment, Phoenix, AZ.

Lewis, D. M, Mitzel, H. C., Green, D. R., & Patz, R. J. (1999). *The bookmark standard setting procedure*. Monterey, CA: McGraw-Hill.

Lindsay, B. G. (1995). *Mixture models: Theory, geometry, and applications*. Hayward, CA: Institute of

Mathematical Statistics.

Linn, R. L. (2000). Assessments and accountability. *Educational Researcher, 29*(2), 4-16.

Linn, R. L. (2003). The bookmark standard setting procedure: strength and weakness. Canada. *Language Learning, 52*(3), 537-64.

National Center for Education Statistics (2008). *The NAEP writing achievement levels.* Retrieved March 16, 2008, from http://nces.ed.gov/nationsreportcard/writing/achieve.asp.

Perie, M. (2005). *Angoff and bookmark methods.* Workshop presented at the annual meeting of the National Council on Measurement in Education, Montreal, Canada, as cited in Karantonis, A., & Sireci, S. G. (2006). The bookmark standard-setting method: A literature review. *Educational Measurement, Issues and Practice, 25*(1), 4-12.

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.

Shrout, P. E. (1988). Measurement reliability and agreement in psychiatry. *Statistical Methods in Medical Research, 7*, 301-317.

Sim, J., & Wright, C. C. (2005). The Kappa statistic in reliability studies: Use, interpretation, and sample size requirements. *Physical Therapy, 85*(3), 257-268.

Skaggs, G., & Tessema, A. (2001). *Item disordinality with the bookmark standard setting procedural.* Paper presented at the 2001 annual meeting of the national council on measurement in education, Seattle, WA.

Swaminathan, H., Hambleton, R. K., & Algina, J. (1974). Reliability of criterion referenced tests: A decision theoretic formulation. *Journal of Educational Measurement, 11*, 263-268.

U. S. Department of Education (1996). *Goals 2000: A progress report.* Washington, DC: The author.

Vermunt, J. K., & Magidson, J. (2005). *Technical guide for Latent GOLD 4.0: Basic and advanced.* Belmont, MA: Statistical Innovations.

Vinovskis, M. A. (1998). *Overseeing the nation's report card: The creation and evolution of the national assessment governing board (NAGB)*. Retrieved March 3, 2008, from http://www.nagb.org/pubs/95222.pdf.

Yin, P., & Schulz, E. M. (2005). *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：2007 年 06 月 12 日
一稿修訂日期：2007 年 11 月 16 日
二稿修訂日期：2008 年 04 月 08 日
三稿修訂日期：2008 年 10 月 14 日
接受刊登日期：2008 年 10 月 15 日

Bulletin of Educational Psychology, 2009, 41(1), 69-90
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

Normal Mixture Model as Convergent Validity Evidence to Bookmark Standard Setting of English Reading and Listening Ability

Yuh-Yin Wu
Department of Psychology
National Taipei University
of Education

Yan-Ming Chen
Department of Education
National Taipei University
of Education

Yuwen Chang
Department of
Children English Education
National Taipei University
of Education

Shu-hui Eileen Chen
Department of Psychology
National Chung Cheng University

Tung-Hsien He
Department of Psychology
National Chung Cheng University

Jyun-Ji Lin
Department of Psychology
National Chung Cheng University

This study investigated convergent validity of the bookmark standard setting method used for English reading and listening ability. The data set was obtained from 2005 Taiwan Assessment of Student Achievement (TASA) data bank. A total of 10101 sixth graders from different areas of Taiwan were cluster sampled and tested by a 40-item scale. The scale was developed through balanced incomplete block design out of 70 items. Thirteen experts formed bookmark standard setting seminar. Among them, 7 were university professors in the English-as-a-Foreign-Language (EFL) field, 3 were professors in measurement, and 3 were elementary school English master teachers. They attained the consensus of cut score Θ=-.57 with 72.7% of students were classified as passed. The result from normal mixture model (Θ=-.40) was consistent with the result from the bookmark standard setting method with classification consistency Kappa=.87, indicating convergent validity evidence. In line with this finding, issues on how to implement bookmark standard setting approach were further explored and discussed.

KEY WORDS: bookmark standard setting method, convergent evidence of validity, English reading and listening ability, normal mixture model, standard setting