國立臺灣師範大學教育心理與輔導學系
教育心理學報，2021，53 卷，2 期，307–334 頁
https://doi.org/10.6251/BEP.202112_53(2).0003

以「補充性表現水平描述輔助自陳式測量構念」之延伸 Angoff 標準設定研究*

謝進昌

國家教育研究院
測驗及評量研究中心

隨著十二年國民基本教育課程推動，其核心重視學生知識、態度、技能與策略等全人培養。為回應這波課程革新對於學生表現影響，遂成立臺灣學生成就長期追蹤評量計畫（TASAL），目的在描述臺灣學生表現及探究影響因子，其本質為標準本位評量。然而，在回顧過去標準設定研究，多以學科認知範疇的標準設定為焦點，較少涉及情意、策略面向。據此，本研究提出補充性表現水平描述（S-PLD）概念，以輔助專家教師使用延伸 Angoff（extended Angoff）標準設定法，進行自陳式測量構念通過分數設定，並透過檢視效度之過程、內部與外部證據，以支持本研究結果合理性。本研究結果顯示標準設定成員對於標準設定過程，多認同其適切性。藉由討論與反思，凝聚出可接受、適當結果，而在檢視成員於各輪次評定結果的穩定與一致性，其誤差也多能在合理範圍內。最後，本研究所設定兩個切截點，也具有一定區別不同層級策略使用者於外在效標（英語文理解）的表現。整體而言，本研究結果是具有過程、內部與外部證據支持，並於文末，提出幾點建議，供未來研究者參考。

關鍵詞：延伸 Angoff 法、自陳式測量、補充性表現水平描述、標準設定、臺灣學生成就長期追蹤評量

* 1. 通訊作者：謝進昌，jin@mail.naer.edu.tw。
2. 作者感謝「國家教育研究院」經費補助與參與標準設定會議之學者專家、教師協助，才得以順利推動本計畫（編號：NAER-107-12-B-1-07-06-1-08）；本文為作者學術觀點，不代表任何正式官方決策，若有任何疑問，請洽作者。

教育部於 108 學年度起，正式逐年推動十二年國民基本教育課程。其理念重視成就每一個孩子，以適性揚才與終身學習為目標，希冀培養學生核心素養，發揮潛在能力；除重視知識培養外，更強調情意、技能、方法與策略學習等發展（「十二年國民基本教育課程綱要總綱」，2014）。在此理念下，為因應學校教與學方式、環境的轉變，以評估國家課程改革對於學生表現影響，國家教育研究院成立臺灣學生成就長期追蹤評量計畫（Taiwan Assessment of Student Achievement: Longitudinal study, TASAL），目的在描述學生表現及探究影響學生成長變化因子（國家教育研究院，無日期）。有鑑於 TASAL 屬性為標準本位評量（standard-based assessment），標準設定遂成為重要任務議題；研究者得以適當方式將 108 課程綱要為基礎之內容標準（content standard），轉換為具體、可操作表現標準（performance standard），進而透過評量與調查工具建構及標準設定程序，進行描述與探究學生成長表現概況。

在回顧過往標準設定研究中，可發現其應用層面多數聚焦於「學科認知表現」，例如，透過專家判讀或統計技術來進行國語文（曾建銘、王暄博，2012a）、英語文（吳毓瑩等人，2009；黃馨瑩等人，2013；謝名娟，2013）、數學（吳宜芳等人，2010；曾芬蘭等人，2017）、社會（曾建銘、王暄博，2012b）、自然（謝進昌等人，2011）與科學探究（林小慧、吳心楷，2019）等學科標準設定。其中，引起研究者關注與發想的議題在於類似的專家判讀標準設定程序，是否可直接運用於「自陳式心理測量構念」（self-reported measures）對於通過分數的決斷？研究者在進一步思考與比較兩者內容屬性差異，可發現執行自陳式測量構念標準設定時，可能會面臨幾項議題。一是自陳式心理測量題項內容較為抽象，不似學科試題來得具體，容易造成標準設定成員在判讀最低水平學生在例如「我喜歡上英文課、我會從上下文或整段內容來理解文章在說甚麼」等題項可能表現時，落於各自表述與判讀；其二是如何對應與校準（align）不同表現水平學生於策略使用、情意態度、與認知表現所設定標準關係，才得以讓決策者所設定的標準能反映出 108 課綱所重視整合認知、情意與技能之整體性素養概念，而非單獨各自設定的向度元素。研究者為企圖回應上述兩個議題，遂提出「以補充性表現水平描述」（supplementary performance level descriptors, S-PLD）概念，來輔助推動以專家判讀為取徑進行自陳式心理測量構念標準設定。

在檢視有關心理測量構念的標準設定，研究者發現過去研究多傾向以填答者整體表現結果，來進行通過分數（passing score）或稱切截、決斷分數（cutoff score）的決斷。就國際大型教育評比計畫而言，國際數學與科學教育成就趨勢調查（Trends in International Mathematics and Science Study, TIMSS）、促進國際閱讀素養研究（Progress in International Reading Literacy Study, PIRLS）等對於學生信心高低分群，是以類似等分（equal division）概念來決斷。假設在 7 個測量學生學科信心題項中，若學生有約略超出一半的 4 題回答非常同意、其餘 3 題回答有點同意，則視為有信心者；以同樣概念進行對稱映照，學生有略超出一半的 4 題回答有點不同意、其餘 3 題回答有點同意，則視為無信心者（Martin et al., 2014, p. 308）。其對於通過分數的決斷，並不關心哪些題項（內容）被回答非常同意或有點不同意，僅考量學生回答總分。這類標準設定策略優點在於直觀、設定所需成本相對較少，可透過常模（或國家）比較，來描述或解釋學生（相對）表現。然而，其通過分數的決斷往往會落於武斷選擇（arbitrary selected），容易忽略每個題項內容差異性與學生期望被達到的表現標準。

為透過專家判讀取徑，來推動自陳式心理測量構念的標準設定，本研究提出以補充性表現水平描述（S-PLD）來輔助專家、教師使用延伸 Angoff 標準設定法（extended Angoff）（Hambleton & Plake, 1995），進行通過分數設定。本研究心理測量構念聚焦「促進英語為外語學習者英語文理解策略的使用」，而補充性 PLD 材料為「學生英語文理解表現水平描述」。其中，提供 S-PLD 優點在於能回應前述議題，不僅能協助專家教師具體化判讀內容與方向，同時能校準不同層級學習者策略使用及其合理對應英語文理解表現水平。整體而言，本研究目的在依據 Kane（1994）建議的標準設定效度評估架構，包含效度的過程、內部與外部證據（procedural, internal, and external evidence for validity），以支持本研究所提出標準設定概念的有效性證據。茲條列本研究待答問題如下：

（一）以補充性表現水平描述（S-PLD）輔助專家教師使用延伸 Angoff 法，進行英語文理解策略使用標準設定效度之過程證據結果為何？

（二）以補充性表現水平描述（S-PLD）輔助專家教師使用延伸 Angoff 法，進行英語文理解策略使用標準設定效度之內部證據結果為何？

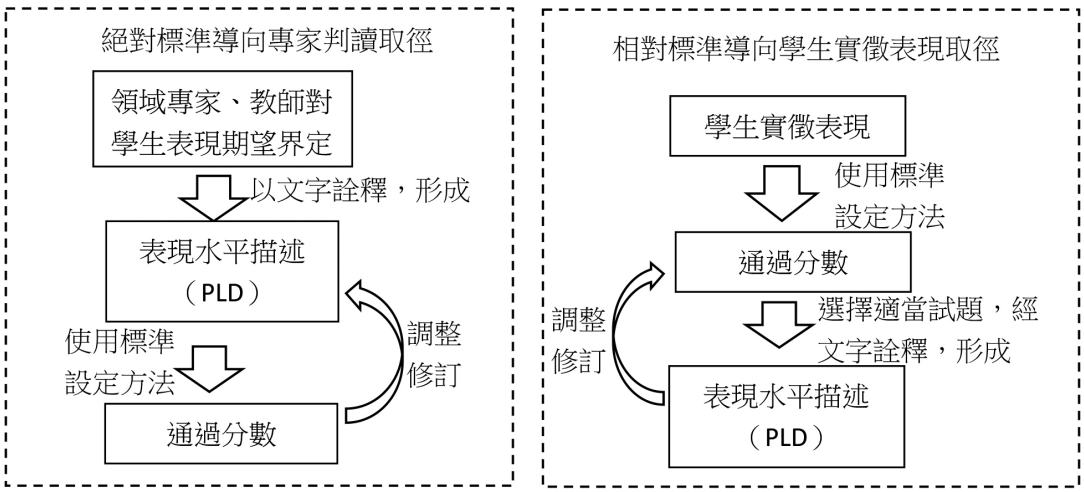
（三）以補充性表現水平描述（S-PLD）輔助專家教師使用延伸 Angoff 法，進行英語文理解策略使用標準設定效度之外部證據結果為何？

文獻探討

（一）絕對與相對標準設定取徑

標準設定（standard setting）係指研究者透過適當方法進行一系列建立標準（standard）或通過分數的過程（Cizek et al., 2004）。研究者在回顧過去標準設定應用研究，發現其發展形態十分多元；然而，在細究各研究者執行標準設定時，其考量的共通元素，大多包含幾個特點，分別是釐清測驗與評量目的、選擇適當的標準設定方法、確定表現標準的分類與命名、撰寫各水平表現標準描述（performance level description, PLD）或稱成就水平描述（achievement level description, ALD）、標準設定成員選擇與訓練、訊息回饋與監控、設定結果有效性的評估等（Cizek & Bunch, 2007; Hambleton, 2001）。隨著研究者目的、使用標準設定方法差異，其本質內涵、執行程序也略具不同，但大致可區分為兩種取徑，如圖 1 所示。一是左側以設定絕對標準為導向的專家判讀取徑、另一是右側以設定相對標準為導向的學生實徵表現取徑。

圖 1
以絕對、相對標準為導向之標準設定取徑



絕對標準（absolute standard）係指該標準反映的是學生所具備的知識、能力（Beuk, 1984）；它往往是透過領域專家、教師（subject matter expert）來詮釋與描述，例如，學生得具備理解文本字面意義的能力，包含大意擷取、指出明顯可見事物特徵等。絕對標準概念重視「政策、專家與教師的期望」，往往會透過一段文字的描述，來形塑政策定義（policy definition）與不同表現水平學生能力標準的說明（Bourque, 2009）。後續，再經選擇適當標準設定方法或技術，由領域專家、教師將文字描述，經檢視評量試題（內容）或學生實質能力，轉換為具體、可量化的通過分數。若就現行大型教育評量與調查研究而言，採用這類取徑進行標準設定者，包含如美國國家教育進展評估（National Assessment of Educational Progress, NAEP）（Sireci et al., 2009）、臺灣學生學習成就評量資料庫（Taiwan Assessment of Student Achievement, TASA）（黃馨瑩等人，2013；謝名娟，2013；謝進昌等人，2011），目的在於反映與回饋學生在國家課程標準的表現。

相對標準（relative standard）係指該標準反映的是相對於其它學生，個別學生的表現結果（Beuk,

1984）。它往往是透過檢視與分析學生實徵表現結果來決定，例如學生是否達到預先設定的通過分數（如百分位數 90、75、50 與 25）。相對標準概念重視「群（個）體相對比較與結果描述」，透過研究者預先決定的通過分數，選擇適當對應各個通過分數點的評量試題，經詮釋題目內容，以作為各層級學生表現結果的詮釋或描述。就現行大型教育評量與調查研究而言，採用這類取徑進行標準設定者，包含有現行國際數學與科學教育成就趨勢調查（TIMSS）、促進國際閱讀素養研究（Progress in International Reading Literacy Study, PIRS）、國際學生成就評比計畫（Programme for International Student Assessment, PISA）等國際大型教育評比調查（Mullis & Prendergast, 2017; Mullis et al., 2016; Organisation for Economic Co-operation and Development [OECD], 2020）。其概念是透過類似量尺定錨（scale anchoring）程序（Beaton & Allen, 1992; Kelly, 1999），經選擇各分數點適當對應題目以描述各層級學生的表現，目的在比較各國家學生表現與政策回饋。

（二）延伸 Angoff 標準設定法

因循不同標準設定取徑發展的脈絡，過去研究者提出許多標準設定方法或技術，大致可區分為以測驗中心（test-centered），例如，Angoff 與相關衍生與修訂技術（Angoff, 1971; Impara & Plake, 1997）與受試者為中心（examinee-centered），例如，對照組法（contrasting group method）（Zieky & Livingston, 1977）、選擇特定百分位數（Mullis & Prendergast, 2017）。

Angoff（1971）法概念在要求標準設定成員針對每一個試題進行判讀，以決定兩水平間臨界受試者（borderline group）或稱該水平最低能力受試者（minimally competent examinee）有多高機率（0-100%），可以正確回應該題目。然而，其原始概念僅適用於 2 元計分，因此，Hambleton 與 Plake（1995）延伸至多元計分題型，提出「延伸 Angoff 法」。其程序要求標準設定成員針對最低能力受試者在每個多元計分試題上，實際可獲得期望分數，而在後續應用中，有些學者會採用 Nassif（1978）所提 yes-no 概念，來簡化判讀複雜性。若將 yes-no 概念加以延伸至延伸 Angoff 中，即要求標準設定成員依照計分準則描述，分別就最低能力受試者是或否能得到 1 分、2 分、3 分等進行判讀。整體而言，本研究所使用標準設定方法概念，為延伸 Angoff 的概念，而其計分依循是、否（yes or no）正確回應該試題的方式，進行計分。

在上述標準設定方法中，可發現學者所提出的標準設定方法，其應用層面大多是以成就、認知題型為導向，有其正確計分方向，待判讀內容與對應表現水平描述較為客觀具體。然而，對於自陳式心理測量構念題型而言，大多是受試者對於自身信念、價值或策略使用的自我認同程度，題項與得分對應較為抽象，例如，我會從上下文或整段內容來理解文章在說甚麼、我喜歡學習英語文等，以非常不同意至非常同意等傾向計分。因此，研究者若直接援用如 Hambleton 與 Plake（1995）延伸 Angoff 法，容易使得成員對於題項的判讀落於自我表述，失去共同參照點。據此，本研究在前述標準設定取徑基礎下，提出以「補充性表現水平標準描述」（S-PLD）以輔助自陳式測量構念進行延伸 Angoff 標準設定法。其概念在於預先提供可輔助標準設定成員具體形塑「該水平最低傾向參與者於該自陳式測量構念實際表現之描述材料」，其優點不僅能具體化各水平學生應具備表現水平描述內容，更能同步銜接、對應其它向度表現標準；後續，研究者再要求標準設定成員針對該水平最低傾向參與者（minimally intentional participant）於每個心理測量構念題項上，是或否展現出的該傾向反應，進行判讀與通過分數估計。

（三）「英語為外語學習者」學習策略使用

探究如何有效的運用學習理解策略以提昇學生英語文學習表現，早已經是過去許多研究者著重的焦點（Ardasheva et al., 2017; Jeon & Yamashita, 2014; Plonsky, 2011），也是 108 課綱重視焦點。學者對於學習策略具體定義與分類，也許略有差異，但其概念大多強調學習者能有意識的使用，以進行有效率的語言學習（Griffiths, 2007）。若就過去統合文獻回顧結果，策略類型大致可區分為多元面向（Barjesteh et al., 2014），而本研究在此僅聚焦記憶、認知、推論連結、理解監控等（謝進昌，2021a）。

就記憶與認知面向，它代表外語學習者透過特定方法、途徑，以促進其對英語文素材記憶、保留與理解。記憶策略係指學生透過有效記憶方式以促進其對於語言學習的理解，它包含如應用聲音、圖像、情境脈絡與字詞連結，產生心理表徵（Gambrell & Bales, 1986）、針對字詞進行有意義歸類、使用某些有意義形態或規則辨識字詞等；認知策略係指學生透過有效的認知理解途徑進行訊息處理與組織，以促進其對於英語文素材學習理解，它包含如摘取大意、作筆記、重點註記、略讀或特定形態規則分析等（Padron & Waxman, 1988）。

就推論連結與理解監控面向，它代表著學生透過連結背景經驗與學習素材，以進行有意義的理解，同時，能監控自身理解概況，以進行調整、檢驗、評估等（Baker & Brown, 1984）。就推論連結而言，其係指學生能透過連結自身背景經驗與知識，經訊息歸納、推論與演繹等，以促進其對於英語文素材學習的理解，它包含如上下文連結、透過脈絡、語言線索或其他顯著特徵進行猜測、肢體語言運用、重複閱讀以尋找關聯等；就理解監控而言，其係指學生對於自身學習歷程、理解過程的覺知，透過檢驗、監控、評估與調整等確認其對於英語文素材理解程度，它包含如監控學習進度概況、重複檢視以確認正確性、調整錯誤、評估與選擇適當素材等（Ardasheva et al., 2017）。

（四）標準設定結果的效度評估

對於效度（validity）的評估，它並非單純全有或全無的狀態，得透過理論與實徵證據的累積，以支持研究者對於測驗分數詮釋、應用的適切性（American Educational Research Association et al., 2014）。若就標準設定情境，過去學者對於效度評估證據來源，大致建議有支持效度的過程、內部與外部證據（procedural, internal and external evidence for validity）（Kane, 1994, 2001）。

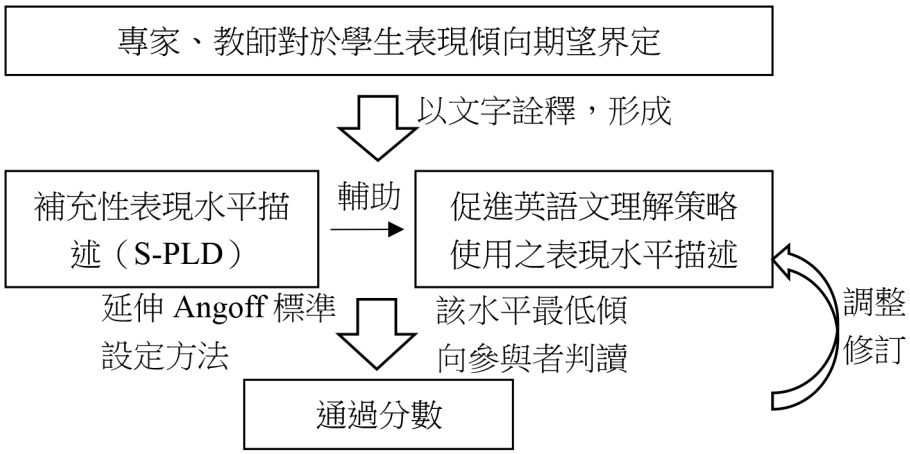
就支持過程有效性證據而言，它重視整個標準設定過程適切與合理性，檢視的內容包含有標準設定方法選擇、不同水平表現標準描述的形成、標準設定成員的招募、訓練、訊息回饋與後續資料分析嚴謹性、甚至是通過分數詮釋與結果影響性等（Kane, 1994, 2001）。內部有效性係指使用標準設定技術時，成員所設定結果的穩定性及一致性，可能涉及成員內、成員間設定一致性，其中，學者對於判讀準則建議存在歧異性，某些學者認為設定的通過分數誤差（standard error of cutoff score）以不超過測量標準誤（standard error of measurement）的四分之一為佳（Jaeger, 1991; Sireci et al., 2009），然而，考量到成員人數多寡會影響誤差可能範圍，Cohen 等人（1999, p. 364）建議以不超過二分之一，而 Kaftandjieva（2010, p. 104）是以折衷的三分之一，作為判斷準則；就外部有效性而言，它係指透過不同標準設定方法結果的比較、或具潛在相關之外在效標，作為設定結果有效性評估的外部支持證據（Kane, 1994, 2001）。

方法

（一）以絕對標準導向專家判讀標準設定程序

本研究目的在設定自陳式心理測量構念標準，採用絕對標準導向專家判讀取徑，如圖 2 所示，經專家、教師預先界定學生表現的期望，以形成促進英語文理解策略使用水平標準描述。在考量到自陳式心理測量構念題項，本身內容較為抽象，不易對應原始水平標準描述，以進行題項判讀，因此，研究者同時搭配使用英語文理解認知表現水平標準描述作為輔助，以提昇專家、教師具體化不同水平學生能達到的表現內容。後續，再透過延伸 Angoff 標準設定方法，以設定出二個切截點，將學生區分為基礎、精熟與高階策略使用者。

圖 2
本研究對於自陳式心理測量構念之專家判讀標準設定取徑



以專家導向進行判讀的標準設定取徑，一般採用的表現水平描述，大多是以原構念（例如，某學科成就表現）進行描述發展，鮮少使用其它輔助材料，然而，考量到本研究關注標的及其發展的描述較為抽象，研究者提出補充性表現水平描述（S-PLD）。它代表著提供能輔助專家、教師具體化各水平學生表現內容或素材，以進行專家判讀標準設定。以本研究為例，如表 1 所示，研究者提供與策略使用具高相關之認知成就表現描述，作為專家進行題項判讀參照點，而其優點在於一能具體化原始描述、二能銜接學生其它向度表現概況。就基礎策略使用者而言，其英語文理解表現對應著 L2 水平學生，能讀懂英語文詞義與直接指出文本內明顯可見訊息，其次，精熟策略使用者的英語文理解表現，對應著 L3 水平學生，能掌握、理解文本整體字面意義（literal meaning）。最後，高階策略使用者英語文理解表現，對應著 L4（以上）水平學生，能理解文本可能隱含意義（implicit meaning）、甚至展現出對於作者文意詮釋與評估表現。後續，研究者採用延伸 Angoff 標準設定方法，要求標準設定成員在檢視表 1 描述後，針對各水平最低傾向參與者（minimally intentional participant）或兩水平間臨界學生可能表現形成具體樣貌並於每個心理測量構念題項上，是或否展現出如 1 分（很少如此）、2 分（有時如此）、3 分（常常如此）等傾向，逐一進行判讀與記錄於如圖 3 之表格，例如，某成員第一輪認為「精熟策略最低使用者」於 M1 題項，至少得「很少」使用該策略，才得視為精熟策略使用者，則在空格處，填入 1；該成員認為「高階策略最低使用者」至少得「常常」使用該策略，則於空格處，填入 3，如此，逐一針對每個策略內各試題，進行第一輪判讀，以利後續通過分數估計。

表 1
本文補充性表現水平描述與促進英語文理解策略使用表現水平描述對應

認知水平	補充性材料：英語文理解表現水平描述	策略使用水平	促進英語文理解策略使用表現水平描述
L 1	本水平學生開始建構學習英語文基本元素，其認知歷程始於學生「記憶、解碼與詞義理解」。學生表現包含： 1. 能聽辨語音（子音、母音）與識讀字母。 2. 能使用字母拼讀規則，聽辨音節結構簡單字詞。 3. 能聽辨、識讀基本字彙、日常且簡單結構句子。	基礎策略使用者	本水平學生僅常常使用極少數且特定（慣用）的記憶、認知策略，而有時會使用部分特定記憶、認知策略，其中，對於多數的推論連結、後設認知策略，則很少涉及，甚至是從來沒有使用某些特定推論連結與後設認知策略。
L 2	本水平強調學生「詞義理解與直接指出明顯可見的訊息」。學生表現包含： 1. 能聽辨、識讀結構較複雜句子，並作出適當回應。 2. 能直接指出明顯可見的重要訊息、或直接連結其關聯，例如，簡單人物、事件、時間、地點或物品的外顯特徵、或彼此直接關聯。		
L 3	本水平強調學生能「直接區分辨別重點與細節、直接連結」，未涉及過多轉折。學生能直接區分辨別、直接連結（結合）文本內重要訊息與細節，並透過自身背景經驗或其它明顯可見線索，直接連結文本內訊息關聯性，藉以理解文本整體字面意義（literal meaning）。學生表現包含： 1. 能直接區分辨別文本內可見的關鍵訊息與細節，或直接連結結合重點，形成整體字面意理解。 2. 能透過自身背景經驗、或外在圖（影）像、圖表、場域、或標題等線索輔助，直接連結文本內完整可見訊息，以直接推論出彼此關聯。	精熟策略使用者	本水平學生常常使用部分（慣用）的記憶、認知策略與極少數推論連結、後設認知策略，而有時、或很少會使用部分特定記憶、認知、推論連結與後設認知策略，其中，對於極少數的推論連結、後設認知策略，則從來沒有使用。
L 4	本水平強調學生能「經訊息統整、歸納，以進行推論」。學生能透過自身背景知識或其它線索，經統整要點與歸納，推論出文本可能隱含意義（implicit meaning）。學生表現包含： 1. 能統整文本概念、論述或個人觀點，推論出文本意涵或主旨、通則、作者立場或意圖。 2. 能區分辨別文本使用語文特徵，如客觀事實與主觀意見、或複雜文本數個訊息相似（異）性等。 3. 能推論出人物隱含意圖、事件因果脈絡、情緒或態度轉折、或基於文意選擇出適當策略等 、或彼此文內未明確說明之關聯。	高階策略使用者	本水平學生常常使用多數的記憶、認知策略與部分的推論連結、後設認知策略，而僅有時會使用極少數（不常見、適用情境不高）的記憶、認知策略、與部分推論連結、後設認知策略。 對於某些極少數且適用情境較低的推論連結、後設認知策略，本水平學生才很少或從來沒有使用。
L 5	本水平強調學生能透過連結自身背景知識，經適當推論文本隱含意義後，進行詮釋與分析、評估。學生表現包含： 1. 能詮釋作者、人物的特質、意圖、情緒或解釋文本主要概念與論點。 2. 能分析、評估概念、觀點立場、正確性、事物發生可能性等。		

註：本表「補充性材料：英語文理解表現水平描述」以標準設定當日呈現為主，然而，相關描述內容後續仍經過數次微調，詳細可見〈第四學習階段英語文素養長期追蹤〉（編號：NAER-107-12-B-1-07-06-1-08），謝進昌，2021a，國家研究院（https://www.naer.edu.tw/bin/home.php）或後續相關發表。

圖 3
本研究延伸 Angoff 標準設定之記錄表格（示意圖）

題號	元素	題項	第一輪		第二輪		第三輪	
			精熟策略使用者最低(L3 最低者）應使用頻率	高階策略使用者最低(L4 最低者）應使用頻率	精熟策略使用者最低(L3 最低者）應使用頻率	高階策略使用者最低(L4 最低者）應使用頻率	精熟策略使用者最低(L3 最低者）應使用頻率	高階策略使用者最低（L4 最低者)應使用頻率
			0 從來沒有	1 很少	2 有時候		3 常常	
M1	圖像化	在學英文單字時，我會試圖把單字的發音和單字代表的圖像連結起來（例如，念 refrigerator（冰箱）時，腦袋會聯想到冰箱的樣子）	1	3				

（二）標準設定成員：人員、經驗與訓練

本研究所邀請參與標準設定成員，總計有 8 名，其背景分佈為女生（7 名）、男生（1 名），分別來自於臺北市（4 名）、新北市（2 名）、高雄市（1 名）、宜蘭（1 名），皆為任教英語文教師，平均教學年資為 18.75 年，而其中 7 名教師皆有參與補救教學專案經驗，能兼顧與熟悉不同層次學生英語文表現概況。

就標準設定相關經驗而言，在推動本研究自陳式心理測量構念標準設定前，此 8 名成員皆曾同步於 2020 年 4 月 8 日參與過英語文理解認知表現之專家判讀標準設定（謝進昌，2021b），不僅熟悉前述表 1 的英語文理解表現水平描述，同時，具備執行專家判讀標準設定方法經驗。就本研究標的訓練而言，除了透過前導資料，包含有研究背景與目的、標準設定技術、策略使用表現水平描述、議程等說明，預先讓標準設定成員熟悉素材外，於 2020 年 5 月 20 日會議當日，研究者以口頭說明前述素材、透過問題與討論、成員試練習標準設定方法等，讓所有成員具備進行本研究標準設定的能力。

（三）標準設定與效度之外部驗證材料：大型教育調查與評量工具

本研究標準設定與效度之外部檢核材料分別為促進英語文理解策略使用之自陳式問卷、與英語文理解表現，而其發展背景皆在於 108 課程綱要脈絡下，調查與評量全臺灣七年級學生表現概況（謝進昌，2021a，2021b）。就抽樣而言，本研究採用兩階段分層叢集抽樣（stratified two-stage cluster sampling），第一階段就各分層內，依學校規模大小進行系統性機率比例抽樣（systematic sampling with probability proportional to sample size）進行學校單位選取，第二階段則是就選取學校內班級為單位，進行隨機選取。本研究正式選取七年級 246 校、計 271 班，總計 7,246 學生，然而，由於每個班級內學生，係依照循環方式，僅接受五科中某兩科測驗，因此，實際參與英語文測驗學生，正式有效樣本為 2,732 名，分別為男生 1,417 名、女生 1,315 名。

本研究英語文策略使用自陳式問卷係依照標準化程序進行開發，經確認目的，發展調查架構、試題題項撰寫與九名專家（包含英語文、閱讀理解、教育心理、測驗與評量專長）修審，分別於 2018 年 5、10 月進行調查工具小規模預試，經確立正式調查工具後，於 2019 年 5 至 6 月進行正式七年級評量調查（謝進昌，2021a）。本研究英語文策略使用自陳式問卷分為記憶（6 題）、認知（6 題）、推論理解（8 題）與理解監控（10 題）等四個策略，而各策略所呼應的架構元素，詳如附件一所示，以記憶策略為例，其下包含透過圖像化、聲韻、脈絡化、搭配詞／同義詞、字根／首／尾、音節等六個元素來促進英語文記憶層面，其下題項則依照此六元素分別書寫。此外，四個策略大致可對應表 1 學生英語文理解表現水平描述，其中，記憶策略的使用呼應 L1 語言基本知識、與 L2 直接擷取明顯可見訊息之表現；認知策略的使用呼應 L3 字面意理解；推論策略使用呼應 L4 隱含意理解；理解監控策略使用呼應 L5 詮釋與評估表現，而其下選項自從來沒有、很少如此、有時候、常

常如此，依續給予 0 至 3 分，所得內部一致性信度係數估計，分別為 0.82、 0.85、 0.89、0.91，經以四因素驗證性因素分析，模型適配指標結果分別為 $\chi^2(399) = 1720.837$（$p < .001$）、CFI = .958、TLI = .954、RMSEA = .035、SRMR = .029，其中，除了卡方值容易因大樣本而顯著外，多符合 Hu 與 Bentler（1999）、McDonald 與 Ho（2002）等人對於合理模式適配指標建議。

本研究作為外部效度評估材料為學生英語文理解表現，其評量架構係與表 1 為一體兩面，主要差別在於 PLD 作為學生表現描述，而評量架構多依此描述，改以條列式指標來呈現，以利作為命題與具體對應依據，例如，直接指出明顯可見的重要訊息；直接推論文本內如人物、事件、主詞與代名詞間關聯性等指標條列，因此，其內涵除了包含理解認知成分，分別為語言基本知識、直接擷取明顯可見訊息、促進字面意理解成分（如直接推論、連結）、促進隱含意理解成分（如統整、推論）、詮釋與評估等，另包含文本類型（text type）、與訊息接收或傳遞媒介（medium）。除了評量語言基本知能的文本外，文本類型主要參考 The Organization for Economic Co-operation and Development（OECD, 2019）對於六種文本類型界定，經依內容成分，區分為 1. 目的在傳遞訊息與通知為主之文本，例如景色描述、介紹指引、社交溝通；2. 目的除了訊息傳遞與知會外，隱含著作者想進一步解釋、探索、說服等，例如，議論文、說明文；3. 為了啟發與娛樂、吸取文學經驗等目的，例如，人物故事與敘事。訊息傳遞媒介區分為閱讀、聽力、與混合文字、圖片、多媒體之線上網路媒介，後續，有別於策略使用調查工具之審核專家，另選擇五名專家（包含英語文、閱讀理解）進行評量架構檢核，詳細發展可見謝進昌（2021a，2021b）。在此架構下，研究者同樣透過上述標準化流（時）程，總計發展 182 題閱讀與聽力試題，以組成 26 卷，每位學生僅接受其中 1 卷，約 28 題，經獲得學生作答反應資料，計算其內部一致性信度係數為 0.846，而各試題加權均方適配度值（information-weighted fit mean square, infit MnSq）界於 0.79 至 1.37 間，多符合 Linacre（2005）所建議 0.5 至 1.5 間適合模式適配度標準。

（四）支持效度證據與資料分析

本研究支持心理測量構念標準設定有效性的證據來源有三，分述如下：

1. 效度之過程證據

本研究提供整個標準設定程序合理性說明、成員對於執行過程內容問卷評估、及成員招募、經驗與訓練描述等，作為效度之過程證據，其中，過程內容評估問卷有三類，主要參考先前 TASA 標準設定程序之評估問卷（黃馨瑩等人，2013；謝進昌等人，2011），經轉化與修訂，以適合本研究主軸與標的，而其題項分別是標準設定成員對於前導資料、會議說明與標準設定方法等理解程度、各輪次回饋訊息理解程度、與整體標準設定結果信心程度等，以五點量表進行評估。

2. 效度之內部證據

本研究檢視標準設定結果誤差的合理性、成員設定一致性等，作為效度之內部證據評估。就誤差評估而言，過去研究者認為影響分類一致性誤差來源，大致有二，分別是通過分數標準誤（standard error, *SE*）與測量標準誤（standard error of the mean, *SEM*）（Kaftandjieva, 2010），就前者，本研究透過 Bootstrapping 方法（Efron, 1981），經 1,000 次反覆計算以擷取通過分數標準誤，而就後者的測量標準誤，研究者透過下列公式 1 進行估計，其中，*SD* 填答標準差（本研究設定為 2）、*rel.* 該調查工具內部一致性信度係數，經計算，本研究策略使用四個向度測量標準誤，分別為 0.84、0.78、0.67、0.59。

$$SEM = SD \times \sqrt{1 - rel.} \qquad \text{（公式 1）}$$

就誤差判讀準則而言，學者曾建議「通過分數標準誤以不超過測量標準誤的某個比重」為原則，然而，由於不同測驗的測量標準誤大多不相同，致使較不適合進行跨測驗（向度）比較，因此，若將此概念經過簡單轉換，其誤差合理性評估指標可以公式 2 來表示，其所得比值已經考量不同測驗量尺，可直接互相比較，而在考量本研究標準設定成員僅有 8 名，人數偏少，因此，判讀準則是以

折衷的 0.33（Kaftandjieva, 2010, p. 104）作為評估與解釋的參照點。

$$\frac{SE(ps)}{SEM} \quad \text{（公式 2）}$$

3. 效度之外部證據

就效度之外部證據而言，其來源有二。一是比較本研究與 TIMSS 及 PIRLS 對於心理構念所使用標準設定方法（本研究稱等分法），兩者所設定結果一致性，其二是透過成員所設定出的兩個切截點，分別檢視不同層級策略使用者於外在效標（即英語文理解）表現的差異程度。

有關學生於英語文理解策略使用能力估計，研究者是以多向度部分計分模式（multidimensional partial credit model）（Masters & Wright, 1997; Yao & Schwarz, 2006）進行加權概似估計值 WLE（weighted likelihood estimation, WLE）（Warm, 1989）計算，再經線性等化（Kolen & Brennan, 2004, p. 31）轉換為平均數 10、標準差 2 之量尺分數（謝進昌，2021a），而對於成員判讀結果的估計，同樣使用上述模式，差別只是在固定前述已知試題參數下，進行通過分數計算。此外，對於學生英語文理解表現能力估計，研究者在經獲得學生作答反應資料後，依照多向度隨機係數多項洛基模式（multidimensional random coefficients multinomial logit model, MRCMLM）（Adams et al., 1997）進行參數估計，首先，在自由估計試題參數估計（設定學生平均表現為 0）後，經建立條件變項（conditioned variables）與固定前述估計試題參數下，透過潛在迴歸分析模式，以估計學生貝氏期望後驗能力值 EAP（Bayesian expected a posterior, EAP）（Bock & Mislevy, 1982），再經線性等化，轉換為平均數 500、標準差 100 之量尺分數，以作為學生分數分析基準。本研究相關估計程序皆是在 R 語言環境，使用 TAM 套件（Robitzsch et al., 2020）進行參數估計。

結果與討論

（一）支持標準設定效度之過程證據

本研究標準設定流程內容如所示，經確目的、建立調查架構、題項編擬與撰寫表現水平描述，決定設定兩個切截點，將學生區分為基礎、精熟與高階策略使用者。在選擇適當標準設定成員後，除於會議前寄送前導資料，以利成員熟悉會議內容外，正式於 2020 年 5 月 20 日召開三輪次標準設定會議。當日會議流程主要分為說明與練習、第一輪標準設定、第一輪訊息回饋與討論、第二輪標準設定、第二輪訊息回饋與討論、第三輪標準設定等，分述如下：

1. 說明與練習

在針對會議目的、流程、標準設定方法等進行說明後，研究者請所有成員就表 1 補充性表現水平描述（S-PLD）與策略使用表現水平描述（PLD）內容，逐一檢視與開始練習延伸 Angoff 法，進行題項判讀及問題、討論。

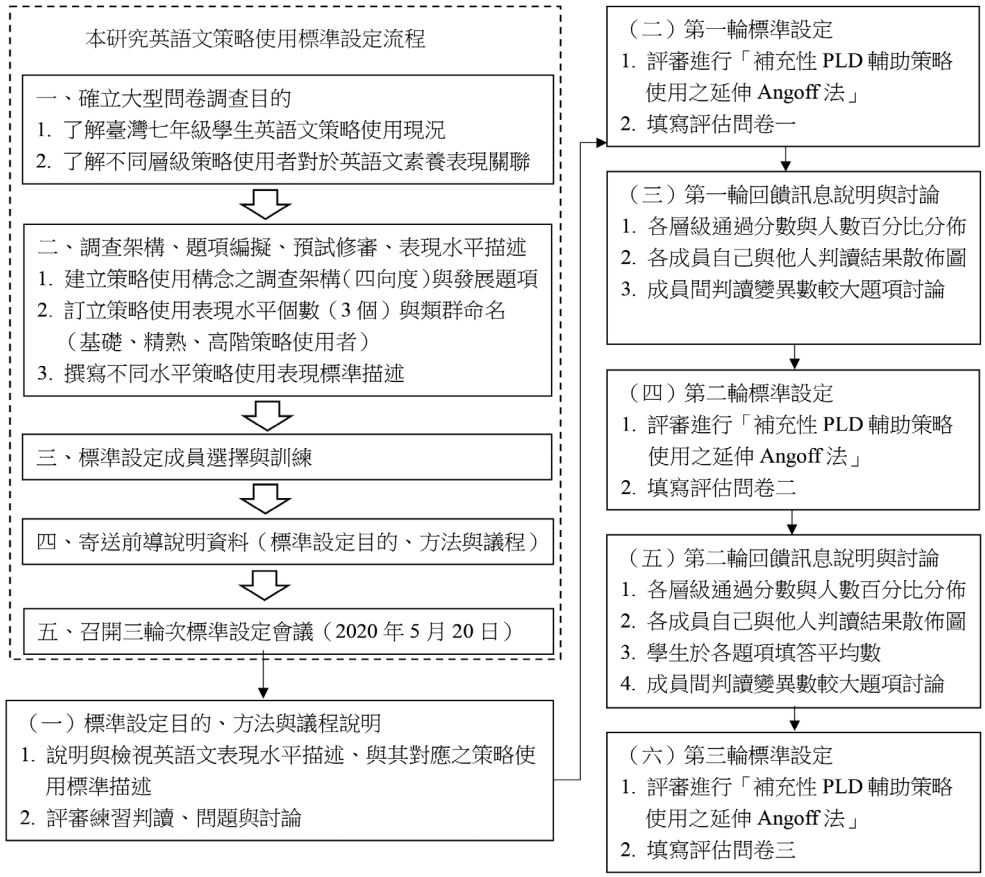
2. 第一輪標準設定

依據補充性表現水平描述（S-PLD）與策略使用表現水平描述（PLD），請標準設定成員就自身專業、教學經驗等，對於精熟水平中最低、高階水平中最低傾向參與者於每個策略使用題項，是／否展現出很少如此、有時如此、常常如此等，逐一進行判讀。待成員完成每一題項的判讀後，研究者回收成員記錄表，並交由兩位分析人員獨立進行分析與交叉比對，以確認分析結果正確性。

在完成第一輪標準設定後，成員會針對評估問卷一，進行填答，其結果如表 2、3 所示。就表 2 而言，成員多同意能理解前導資料說明、會議目的與流程、延伸 Angoff 標準設定方法的運用、促進英語文理解策略使用表現水平描述等，同時，8 名成員也都傾向同意先前曾執行認知層面標準設定經驗，是有助於瞭解現行策略使用標準設定，此外，對於本研究額外提供補充性英語文理解表現

水平描述之助益性評估，如表 3 所示，成員對於只有提供策略使用表現水平描述助益性認同，為 2 名表示非常同意、4 名表示有點同意、2 名表示普通，然而，若額外增加英語文理解表現水平描述，其認同度變成 3 名表示非常同意、5 名表示有點同意，顯示出本研究提出的補充性表現水平描述（S-PLD）以輔助延伸 Angoff 進行標準設定，有其正向助益性。

圖 4
本研究標準設定流程



3. 檢視第一輪回饋訊息與討論

經分析每位成員判讀結果，研究者提供每位成員自己與相對他人判讀分數的散佈圖、兩個切截通過分數及其對應基礎、精熟與高階策略使用者人數百分比分佈等屬於 Cizek 與 Bunch（2007）建議常模參照（normative feedback）與影響性回饋訊息（impact feedback），讓成員了解自己與相對他人設定結果及整體設定通過分數可能產生的結果影響，後續，研究者逐一針對成員間判讀變異數較大的題項，進行公開性討論與文字記錄，以期成員充分了解自身與他人判讀不一致原因。

4. 第二輪標準設定

在經討論後，研究者會要求成員獨自進行第二輪標準設定，其執行程序與內容大致雷同於第一輪，待完成後，再進行評估問卷二的填寫，其結果如表 4 所示，成員對於第一輪回饋訊息，大多同意能理解其他成員（與自己）判定決斷分數散佈圖、「基礎、精熟、高階」策略使用層級通過人數

百分比等意義，同時，也都認同判定不一致題項討論是有助於瞭解其它成員判讀想法。

5. 檢視第二輪回饋訊息與討論

研究者第二輪提供的回饋訊息多雷同第一輪訊息類型，但新增學生於每個題項填答平均頻率數據，以便成員了解學生對於每個策略題項的平均使用概況，如表 4 所示，多數成員也多同意能理解此事實性回饋訊息（reality feedback）意義。

6. 進行第三輪標準設定

第三輪標準設定程序是雷同第二輪程序，而成員對於最後的標準設定結果，如表 4 所示，分別有 4 名成員非常同意、3 名有點同意與 1 名普通同意設定結果的有效與合理性，整體而言，多數成員均對於設定結果具有相當信心。

表 2
評估問卷一：前導資料、會議說明、標準設定方法等可理解性評估

評估問卷題項	非常同意	有點同意	普通	有點不同意	非常不同意
會議前收到的「前導資料」能幫助我瞭解本次會議目的與自身扮演的角色：	6 (75.0%)	2 (25.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
本日「標準設定說明」能幫助我了解本次會議內容及流程：	8 (100.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
我瞭解本次標準設定「會議目的與流程」：	7 (87.5%)	1 (12.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
我瞭解英語文「策略使用各水平表現標準描述（PLD）」內容：	4 (50.0%)	4 (50.0%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
我瞭解如何執行「extended Angoff 標準設定法」：	4 (50.0%)	3 (37.5%)	1 (12.5%)	0 (0.0%)	0 (0.0%)
先前曾執行過「認知層面標準設定經驗」有助於我瞭解現行「策略使用層面」標準設定的進行：	3 (37.5%)	5 (62.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

表 3
評估問卷一：補充性英語文理解表現水平描述之助益性評估

評估問卷題項	非常同意	有點同意	普通	有點不同意	非常不同意
基礎／精熟策略使用者之表現標準描述（PLD）有助於我判別精熟策略使用之最低能力者應具備的頻率：	2 (25.0%)	4 (50.0%)	2 (25.0%)	0 (0.0%)	0 (0.0%)
精熟／高階策略使用者之表現標準描述（PLD）有助於我判別高階策略使用之最低能力者應具備的頻率：	2 (25.0%)	4 (50.0%)	2 (25.0%)	0 (0.0%)	0 (0.0%)
額外提供基礎／精熟策略使用者所對應之英語文理解表現標準描述（界於 L2 ／ L3 層級）有助於我掌握該類群臨界學生的策略使用概況：	3 (37.5%)	5 (62.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
額外提供精熟／高階策略使用者所對應之英語文理解表現標準描述（界於 L3 ／ L4 層級）有助於我掌握該類群臨界學生的策略使用概況：	3 (37.5%)	5 (62.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)

註：界於 L2 ／ L3、L3 ／ L4 層級臨界學生，即是「該水平 L3、L4 最低能力者」。

表 4
評估問卷二、三：成員對回饋訊息可理解程度與整體標準設定結果信心

評估問卷題項	非常同意	有點同意	普通	有點不同意	非常不同意
我瞭解「其他成員（與自己）判定決斷分數散佈圖」的說明：	7 (87.5%)	1 (12.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
我瞭解「基礎、精熟、高階」等策略使用層級通過人數百分比」的說明：	7 (87.5%)	1 (12.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
透過「判定不一致策略使用題項之討論」，我能瞭解其他成員的想法：	7 (87.5%)	1 (12.5%)	0 (0.0%)	0 (0.0%)	0 (0.0%)
我瞭解「學生填答平均頻率數據」的說明：	7 (87.5%)	0 (0.0%)	1 (12.5%)	0 (0.0%)	0 (0.0%)
本次標準設定會議能設定出適合臺灣七年級各層級英語文策略使用學生表現決斷分數：	4 (50.0%)	3 (37.5%)	1 (12.5%)	0 (0.0%)	0 (0.0%)

（二）支持效度之內部證據

1. 標準設定技術內一致性

本概念在於假設若不斷重複執行同一個標準設定方法，可以支持其結果穩定與一致性的證據，而本研究實際是透過 Bootstrapping 方法，經 1,000 次反覆計算，以仿照前述重複執行概念，經擷取的通過分數及其標準誤，如表 5 所示，就記憶策略而言，在第一輪次，標準設定成員所設定的兩個切截點，其通過分數分別是 9.39 分（*SEM* = 0.24）、11.29（*SEM* = 0.30），可發現隨著輪次增加，多能獲得較穩定且相對較小的通過分數標準誤，其中，對於記憶與推論連結策略，成員大致是在第二輪，達到最小的通過分數標準誤，而認知與理解監控策略，成員則是在第三輪，達到最小的通過分數標準誤。

表 5
本研究各輪次通過分數及其標準誤

向度	輪次	通過分數（標準誤）	
		第一切截點	第二切截點
記憶策略	1	9.39 (0.24)	11.29 (0.30)
	2	9.67 (0.07)	11.74 (0.18)
	3	9.64 (0.08)	11.51 (0.28)
認知策略	1	8.70 (0.38)	11.10 (0.32)
	2	8.91 (0.18)	10.96 (0.24)
	3	8.95 (0.11)	10.83 (0.13)
推論連結策略	1	9.15 (0.41)	11.33 (0.28)
	2	9.56 (0.19)	11.81(0.13)
	3	9.50 (0.21)	11.74 (0.28)
理解監控策略	1	8.01 (0.42)	10.54 (0.40)
	2	8.61 (0.20)	10.75 (0.24)
	3	8.66 (0.14)	10.78 (0.22)

在比較與檢視跨向度設定結果時，研究者經計算通過分數標準誤相對於測量標準誤的比值後，其分類一致性誤差如表 6 所示，可發現不分輪次，各向度誤差範圍分別是界於 0.08 至 0.36（記憶策略）、0.14 至 0.49（認知策略）、0.19 至 0.61（推論連結）、0.24 至 0.72（理解監控策略），似乎顯示認知層次愈高的策略，誤差區間愈大，反映出成員間判讀愈傾向不一致，推測其原因可能在於認知層次愈高的策略題項，其內容抽象程度愈大，成員較不容易想像或連結學生可能的表現，例如，理解監控策略的 S1「我會留意自己英語文學習的進步情況」，會比記憶策略 M1「在學英文單字時，我會試圖把單字的發音和單字代表的圖像連結起來」，來得抽象。此外，若以各向度獲得最小誤差輪次，作為最後的通過分數（即記憶、推論連結策略的第二輪次、認知與理解監控策略的第三輪次），可發現僅有理解監控策略第二切截點的誤差 0.37，未符合 Kaftandjieva（2010, p. 104）所建議 0.33 折衷準則。

表 6
本研究各輪次標準設定分類一致性誤差

向度	輪次	通過分數標準誤／測量標準誤	
		第一切截點	第二切截點
記憶策略	1	0.29	0.36
	2	**0.08**	**0.21**
	3	0.10	0.33
認知策略	1	0.49	0.41
	2	0.23	0.31
	3	**0.14**	**0.17**
推論連結策略	1	0.61	0.42
	2	**0.28**	**0.19**
	3	0.31	0.42
理解監控策略	1	0.72	0.68
	2	0.34	0.41
	3	**0.24**	**0.37**

註：反黑底線為各向度獲得最小誤差輪次，為最後通過分數之參照。

2. 標準設定成員內設定結果一致性

此證據概念重視標準設定成員於各輪次內與各輪次間，本身評定結果的穩定與一致程度，就成員於各輪次內設定結果而言，研究者經逐一計算該輪次設定平均數，經加或減 1.96 個成員間設定標準差，作為極端值檢視，可發現絕大多數成員設定結果，皆可落於此區間範圍內，顯示各輪次內，大多數成員並未有不合理的極端判讀；就成員於各輪次間設定結果而言，如表 7 至 10 所示，研究者經逐一計算成員於各輪次設定前、後通過分數之差異絕對值的平均數及其標準誤、與該標準誤對照測量標準誤之比值，就表 7 記憶策略而言，顯示成員內第 2 輪減第 1 輪次的誤差比值，分別為 0.18（第一切截點）、0.13（第二切截點）、成員內第 3 輪減第 2 輪次的誤差比值，分別為 0.08（第一切截點）、0.36（第二切截點），僅有 0.36 約略超過 Kaftandjieva（2010, p. 104）的 0.33 折衷準則建議。另一方面，研究者以此概念分別檢視其餘三向度策略的誤差比值區間，分別是 0.16—0.35（認知策略）、0.16—0.32（推論連結策略）、0.28—0.61（理解監控策略），可發現除了 0.35（認知策略）與 0.37、0.42、0.61（理解監控策略）外，其餘皆能符合 Kaftandjieva 建議準則。整體而言，成員於各輪次前、後間判讀，尚具有一致性。

表 7
標準設定成員於記憶策略各輪次內與各輪次間判讀結果

向度	成員	第 1 輪	第 2 輪	第 3 輪	第 1 輪	第 2 輪	第 3 輪
		第一切截點			第二切截點		
記憶策略	1	9.25	9.77	9.50	12.02	12.70	10.36
	2	9.77	9.50	10.05	12.02	11.51	12.70
	3	10.05	9.77	9.77	11.08	12.02	12.02
	4	9.50	9.77	9.50	11.08	11.51	11.51
	5	9.25	9.77	9.77	10.70	11.08	12.02
	6	10.36	9.77	9.77	12.70	12.02	12.02
	7	8.75	9.25	9.25	10.05	11.08	11.08
	8	8.20	9.77	9.50	10.70	12.02	10.36
成員內第 2 輪減第 1 輪	差異絕對值平均數		0.57			0.75	
	差異絕對值平均數之標準誤		0.15			0.11	
	差異絕對值平均數之標準誤／測量標準誤		0.18			0.13	
成員內第 3 輪減第 2 輪	差異絕對值平均數		0.17			0.77	
	差異絕對值平均數之標準誤		0.07			0.30	
	差異絕對值平均數之標準誤／測量標準誤		0.08			0.36	

表 8
標準設定成員於認知策略各輪次內與各輪次間判讀結果

向度	成員	第 1 輪	第 2 輪	第 3 輪	第 1 輪	第 2 輪	第 3 輪
		第一切截點			第二切截點		
認知策略	1	8.54	9.43	8.84	10.87	11.74	10.10
	2	9.13	8.54	9.13	11.74	10.87	11.29
	3	7.92	8.24	8.84	10.87	10.10	10.87
記憶策略	4	8.84	9.13	9.43	10.87	10.87	11.29
	5	9.13	9.13	9.13	11.29	10.47	10.87
	6	10.87	8.84	9.13	12.89	10.87	10.87
	7	7.57	8.24	8.54	9.76	10.47	10.47
	8	7.57	9.76	8.54	10.47	12.25	10.87
成員內第 2 輪減第 1 輪	差異絕對值平均數		0.87			0.98	
	差異絕對值平均數之標準誤		0.27			0.21	
	差異絕對值平均數之標準誤／測量標準誤		0.35			0.27	
成員內第 3 輪減第 2 輪	差異絕對值平均數		0.49			0.63	
	差異絕對值平均數之標準誤		0.13			0.20	
	差異絕對值平均數之標準誤／測量標準誤		0.16			0.26	

表 9
標準設定成員於推論連結策略各輪次內與各輪次間判讀結果

向度	成員	第 1 輪	第 2 輪	第 3 輪	第 1 輪	第 2 輪	第 3 輪
		第一切截點			第二切截點		
推論連結策略	1	7.93	9.68	8.77	11.11	12.19	10.80
	2	8.98	9.20	9.68	11.11	11.78	12.19
	3	10.22	8.98	9.94	12.19	11.11	12.74
	4	10.22	10.51	10.22	11.78	12.19	12.19
	5	9.43	10.22	9.94	11.78	12.19	11.78
	6	10.80	9.68	9.94	12.19	11.78	12.19
	7	7.93	8.77	8.98	9.68	11.43	11.78
	8	7.70	9.43	8.56	10.80	11.78	10.22
成員內第 2 輪減第 1 輪	差異絕對值平均數		1.00			0.85	
	差異絕對值平均數之標準誤		0.19			0.16	
	差異絕對值平均數之標準誤／測量標準誤		0.29			0.23	
成員內第 3 輪減第 2 輪	差異絕對值平均數		0.53			0.77	
	差異絕對值平均數之標準誤		0.11			0.22	
	差異絕對值平均數之標準誤／測量標準誤		0.16			0.32	

表 10
標準設定成員於理解監控策略各輪次內與各輪次間判讀結果

向度	成員	第 1 輪	第 2 輪	第 3 輪	第 1 輪	第 2 輪	第 3 輪
		第一切截點			第二切截點		
理解監控策略	1	6.03	8.99	8.46	9.37	11.26	10.28
	2	8.99	8.63	9.18	11.26	10.77	11.52
	3	7.55	7.55	8.46	10.77	9.37	10.52
	4	8.63	8.63	8.99	10.28	10.28	11.26
	5	7.34	8.99	8.99	9.37	11.02	11.02
	6	10.28	8.81	8.81	12.94	11.02	11.02
	7	7.74	7.93	8.46	9.58	10.52	11.02
	8	7.55	9.37	7.93	10.77	11.79	9.58
成員內第 2 輪減第 1 輪	差異絕對值平均數		1.06			1.16	
	差異絕對值平均數之標準誤		0.36			0.22	
	差異絕對值平均數之標準誤／測量標準誤		0.61			0.37	
成員內第 3 輪減第 2 輪	差異絕對值平均數		0.54			0.82	
	差異絕對值平均數之標準誤		0.17			0.25	
	差異絕對值平均數之標準誤／測量標準誤		0.28			0.42	

3. 成員間判讀不一致成因分析

在第一輪次判讀與回饋訊息說明後，各成員會針對各向度中判讀變異數較大的題項，進行討論與意見分享，而從成員討論的內容分析，可發現成員間判讀不一致來源，大致可歸納為成員對於題項判讀參照點的不同，而其參照點來源大致包含有「想像學生實際可能表現、銜接表現水平描述、反映專家教師期望」等。

就記憶策略使用的第 2 題項（M2）——聲韻而言，成員判讀不一致原因，其一在於成員對於學生實際可能運用該策略想像的差異，就誠如成員 1 表示「實際上我沒有去看長母音這件事。我直接看就是 old、cold、hold，我的學生也不知道長、短母音，國中端的學生很容易能做類比和串聯」，他認為國中學生使用發音類比與連結，是簡單的，因此，給予較高的判讀分數，而成員 3 也有類似看法，表示「因為 7 年級的學生他們本來就很會聽，所以他們很自然地唸出來」。然而，成員 5、8 認為學生多半不了解長短母音，較難使用該策略，而給予較低的判讀分數，誠如成員 8 表示「我對於學生的理解，就是說我的學生在記單字這種，子音、母音可能有一半學生不知道，但卻知道這個單字，我的判斷是他們對於這個音是子音、母音、長母音、短母音不知道，一樣可以記單字」，而成員 5 同樣表示「我的學生也是這個狀況，他其實真的不知道長、短、子、母音，但他們還是可以把這個單字背起來」。

就前述同樣 M2 題項而言，有別於著眼學生實際可能運用策略的想像，部分成員對於題項的判讀，是從銜接認知表現水平描述著手，其參照來源是完全不同於前述成員，此為形成成員判讀不一致原因二。如同委員 2 表示「基礎策略使用者，學生能了解字母拼讀規則，能夠拼音節結構較為複雜之字詞，最後依它上面的水平標準描述來判斷的話，像 M2 ——聲韻與 M6 ——音節，都大致要會，因為那是基本的」；另一方面，就推論連結向度 I7 題項 - 使用背景知識／經驗進行推論連結而言，委員 2 同樣表示「I7 的部份給的分數也是高的，在那個表現標準描述，精熟策略使用者的部份，有說到他有能力去透過自身背景去摘要出，的確像我們的學生是需要這個能力」。

有別於依循前述參照點，部分委員對於題項判讀，是從自身（專家教師）期望著眼，進而形成判讀不一致原因三。就 I7 題項而言，誠如成員 4 表示「學生需要去做推論連結的這個基礎上面去思考的話，那我的設定是不管是初階、或是中高階的學生，如果你今天受限於單字、文法，受限於其他的干擾因素下，能夠幫助自己的應該是要會去串聯，所以往上往下看應該是他要去解決的能力，應該是能去做得到的，所以我給的分數是比較高的」，然而，秉持類似從專家教師自身期望進行判讀原則但不同看法者，成員 1 表示「就回應背景知識，我覺得閱讀的背景知識真的是和他的家庭背景有關，我在精熟這邊 I7 的部份我是給 0，因為我覺得他就沒有那個 expose 的機會，那我覺得如果這個變成必要的，會太為難了」，因此，給予較低的判讀分數。

整體而言，成員對於題項判讀不一致成因，多是對於題項判讀參照點的差異，某些成員參照來源是「想像學生實際可能表現」，某些成員則是從「銜接表現水平描述、或反映專家教師期望」著手，更甚者，即使在相同參照點內，不同成員又可能秉持不同觀點，這些成因皆可能形成成員判讀差異。

（三）支持效度之外部證據

有關支持外部效度證據，來源有二，一是就本研究設定與 TIMSS 及 PIRLS 所使用設定概念（本研究稱為等分法）（Martin et al., 2014, p. 308），進行結果比較；另一則就本研究所設定不同層級學生於外在效標表現進行檢視。

就應用 TIMSS 及 PIRLS 對於心理構念標準設定概念，在記憶策略 6 個題項中，如果學生有一半的 3 題回答常常如此、其餘 3 題回答有時候，則視為高階策略使用者，以同樣概念對稱映照，學生有一半的 3 題回答很少如此、其餘 3 題回答有時候，則視為基礎策略使用者，經計算，其通過原始分數分別為 $3\times3+3\times2=15$、$3\times2+3\times1=9$，經對照量尺分數，可得使用等分法於記憶策略之通過分數分別為 9.50、11.50，而以同樣概念進行設定，可得其餘向度標準如表 11 所示。研究者經比較本研究設定結果與均分法設定結果，可發現兩者分類一致性最低為認知策略的 81.2%、最高為記憶策略的 90.25%，整體而言，兩種方法分類結果是具有相當一致性。

就本研究成員所設定出的兩個切截點，檢視不同層級策略使用者是否同樣能反映在外部效標表現方面，本研究經以分類一致性誤差最小的第二輪（記憶、推論連結策略）與第三輪（認知、理解監控策略）設定結果，作為通過分數，將學生區分出基礎、精熟與高階策略使用者，進行分別檢視不同層級策略使用者於英語文理解表現的差異程度，其結果如表 12 所示，就記憶策略而言，基礎層級策略使用者的英語文理解平均表現為 454.26 分、精熟層級策略使用者平均表現為 506.86 分、高階策略使用者平均表現為 542.87 分，大致顯示出愈高層級策略使用者，其認知表現愈佳，整體單因子變異數分析 F 統計量為 158.99（$p < .001$）、Partial eta square 為 .104，約為 Cohen（1988）建議界於中等至大效果量。以此概念分別檢視其餘三向度的分析結果，可發現 Partial eta square 皆具備 Cohen 建議大效果量程度，整體而言，本研究所設定通過分數，能有效區別出不同英語文理解表現。

表 11
本研究與國際大型評量均分法之設定結果一致性

方法	記憶策略（題數 = 6）	認知策略（題數 = 6）	推論連結策略（題數 = 8）	理解監控策略（題數 = 10）
本研究兩個切截分數	9.67、11.74	8.95、10.83	9.56、11.81	8.66、10.78
TIMSS 與 PIRLS 均分法	9.50、11.50	9.43、11.73	9.19、11.42	9.17、11.51
分類一致性（%）	90.25	81.20	82.52	87.56

表 12
不同層級策略使用者於英語文理解表現之單因子變異數分析

向度	層級	人數	平均數	標準差	F 值	Partial eta square
記憶策略	基礎	1010	454.26	95.39	158.99 ($p < .001$)	.104
	精熟	1286	506.86	95.21		
	高階	432	542.87	89.95		
認知策略	基礎	703	427.38	78.30	335.37 ($p < .001$)	.198
	精熟	1144	493.88	94.03		
	高階	881	544.49	91.59		
推論連結策略	基礎	947	428.41	75.00	434.23 ($p < .001$)	.242
	精熟	1385	518.57	93.95		
	高階	396	558.64	87.96		
理解監控策略	基礎	593	416.98	70.34	381.50 ($p < .001$)	.219
	精熟	1350	493.69	92.49		
	高階	783	549.61	92.67		

註：Cohen（1988）建議 Partial eta square 0.01 為小、0.06 中等、0.14 為大效果量。引自 "*Statistical power analysis for the behavioral sciences* (2nd ed., pp. 284–287)", by J. Cohen, 1988 (https://www.utstat.toronto.edu/~brunner/oldclass/378f16/readings/CohenPower.pdf). Copyright © 2021 by Lawrence Erlbaum Associates.

結論與建議

本研究目的在透過補充性表現水平描述（S-PLD）以輔助專家教師使用「延伸 Angoff」標準設定法，進行心理測量構念通過分數設定，其中，提供 S-PLD 優點在於能促進專家教師具體化判讀內容與方向，同時能銜結不同層級學習策略使用及其合理對應英語文理解認知表現，而本研究是從過程、內部與外部效度等面向，提供證據以支持結果合理性。茲綜整幾項結論與建議如下：

（一）結論

本研究是以專家導向進行判讀的標準設定取徑，其執行元素大多符合 Cizek 與 Bunch（2007）、Hambleton（2001）建議，將標準設定概念元素融入問卷調查編製過程中，其中，就整體過程面向，經成員評估，大多同意能瞭解本研究所提供前導資料說明、會議流程、標準設定方法使用、回饋訊息、補充性表現水平描述等，同時，成員對於最後所設定的通過分數，也具有相當的信心。此外，標準設定成員的招募，人數雖然僅有 8 人，但皆具備相關標準設定經驗與英語文專業，透過訓練與導引，更能融入整個標準設定過程與討論。整體而言，成員對於標準設定過程安排，多認同其適切性，而藉由討論、分享與反思、調整，凝聚出可接受、適當結果，是具有良好的過程效度。

就檢視支持內部效度證據而言，各向度於各輪次的通過分數標準誤，大多隨著輪次增加，愈趨小，顯示成員間判讀愈趨一致，其中，記憶與推論連結策略是於第二輪具有最小通過分數標準誤，而認知與理解監控策略是於第三輪具有最小通過分數標準誤，其中，除了理解監控策略第二切截點的分類一致性誤差 0.37，稍微超過學者所建議 0.33 折衷準則外，其餘皆符合準則；此外，就跨向度比較而言，結果顯示認知層次愈高的策略，分類一致性誤差區間愈大，反映出成員間判讀傾向愈不一致，推測原因在於認知層次較高的策略，其題項內容較抽象，成員相對不易想像與連結學生可能表現。另一方面，研究者在檢視標準設定成員於各輪次內與各輪次間，本身評定結果的穩定與一致性時，發現各輪次內，成員並未有不合理的極端值，而成員於各輪次前、後間判讀，除了少數誤差較大，尤其是理解監控策略外，其餘誤差尚能符合準則，顯示成員於輪次前、後判讀變異，大多仍維持一定穩定性。最後，在分析成員對於題項判讀不一致成因，就討論內容分析結果，顯示可能在於成員對於題項判讀參照點來源的差異，某些成員的判讀，是參照「想像學生實際可能表現」，某些成員則是從「銜接表現水平描述、或反映專家教師期望」著手，更甚者，即使對應相同參照來源，不同成員又可能秉持不同觀點，而這些成因皆可能形成成員判讀差異，建議未來可考量在導引說明時，就預先就上述成因，凝聚成員共識，可減少磨合、討論時間。

就檢視外部效度證據而言，本研究在檢視不同方法設定結果、及不同層級策略使用者於英語文理解表現差異，結果顯示本研究所設定結果與 TIMSS 及 PIRLS 均分法所設定結果，具有相當一致性，此外，本研究所設定兩個切截點，所產生不同層級策略使用者，能有效區別外在效標的表現，愈高層級策略使用者，其英語文理解表現愈佳，其中，成員於記憶策略所設定兩個切截點，是具有 Cohen（1988）所建議界於中等至大程度差異效果量，而成員於認知、推論連結與理解監控策略所設定出兩個切截點，則是具有大程度差異效果量。整體而言，本研究標準設定結果是具有一定程度外部效度。

（二）建議

成員在連結題項與該水平最低表現（或傾向）學生可能表現，以進行標準設定時，表現水平描述（PLD）往往扮演著一個很重要的角色，它可以凝聚成員的共識與確立判讀方向，然而，在執行心理測量構念的標準設定中，由於該構念所形成題項與內容描述，多較抽象，致使過往研究較常採用以受試者為中心標準設定方法，因此，本研究建議可提供以其它可補充或協助成員具體化不同水平學生表現內容的補充性表現水平描述（S-PLD），來進行以測驗為中心標準設定，經本研究分析，結果顯示此 S-PLD 是具有其助益性，建議未來研究者可嘗試將此概念應用至設定學生情意向度標準，例如，學習動機、態度等，以發揮出採用專家判讀所設定標準與跨向度銜接優勢。此外，十二年國民基本教育課程重視學生認知、策略與情意等全人發展，而國內對於促進中文閱讀理解策略教學已經推動許久且有穩定政策、大型評量的投入（柯華葳，2020），反觀英語文領域，教師教學傾向重視知識內容教導（如文法、句構），似乎相對缺乏系統性、有效的策略教學政策與大型評量，而透過本文所發展英語文認知理解、策略使用之表現水平描述及其間彼此銜接對應，將有助於教師得適性依照學生認知理解水平，教導其適合的理解策略或決策者發展相關的政策方案，建議可進一步推廣與規劃。

本研究透過 S-PLD 輔助專家教師判讀所設定結果，雖然與 TIMSS 及 PIRLS 均分法所設定結果，具有相當一致性，然而，細究方法的本質與內涵，兩者是具有不同的優缺點，前者優點在透過專家教師，逐題檢視題項，其結果能實質反映出成員對於學生期望表現內容，但缺點在於設定所需時間與人力成本較高；後者優點在於設定成本較低與直觀，然而，其缺點是忽略不同題項內容與學生表現期望，單純僅考量填答總分。整體而言，建議未來研究者可依自身需求與現況，選擇適當方法進行心理測量構念之標準設定。

在建立支持標準設定效度的證據中，外部的證據較難收集，一般多會以不同標準設定方法、不同外在效標等，所產生的結果，進行交叉檢核，然而，究其本質，其資料來源大多屬於同時期、橫斷面的證據，鮮少觸及可預測、或連結學生未來可能表現者，據此，在考量本研究屬於長期追蹤大型評量研究，建議未來可針對當同批學生升至八年級、九年級時，以檢視、分析不同層級策略使用者是否同樣具有差異的成長表現，作為具有可預測性之外部效度證據。

在檢視標準設定效度之內部證據中，「誤差」大多是研究者關注標的，而依據學者建議，多是以「通過分數標準誤以不超過測量標準誤的某個比重」為原則，然而，因為不同測驗的測量標準誤，多不相同，致使直接以「通過分數標準誤」進行跨測驗（或向度）比較或直接對照學者建議準則，較容易被質疑未考量測量標準誤，例如，黃馨瑩等人（2013）。因此，Kaftandjieva（2010）建議可以「通過分數標準誤對照測量標準誤的比值」所計算出的誤差比值，來進行分類一致性評估，其優點在於考量了不同測驗量尺（或測量標準誤），可進行跨向度（測驗、跨標準設定研究）比較，更方便直接對照學者建議的準則。

參考文獻

十二年國民基本教育課程綱要總綱（2014 年 11 月）。[Curriculum Guidelines of 12-Year Basic Education: General Guidelines. (2014, November).]

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。[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.] https://doi.org/10.7108/PT.201003.0001

吳毓瑩、陳彥名、張郁雯、陳淑惠、何東憲、林俊吉（2009）：〈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〉。《教育心理學報》，*41*（1），69–89。[Wu, Y.-Y., Chen, Y.-M., Chang, Y.-W., Chen, S.-H. E., He, T.-H., & Lin, J.-J. (2009). Normal mixture model as convergent validity evidence to bookmark standard setting of English reading and listening ability. *Bulletin of Educational Psychology, 41*(1), 69–89.] https://doi.org/10.6251/BEP.20081015

林小慧、吳心楷（2019）：〈科學探究能力評量之標準設定與其效度檢核〉。《教育心理學報》，*50*（3），473–502。[Lin, H.-H., & Wu, H.-K. (2019). Validating the standard setting on multimedia-based assessment of scientific inquiry abilities. *Bulletin of Educational Psychology, 50*(3), 473–502.] https://doi.org/10.6251/BEP.201903_50(3).0005

柯華葳（2020）：〈臺灣閱讀策略教學政策與執行〉。《教育科學研究期刊》，*65*（1），93–114。[Ko, H.-W. (2020). Reading policy and reading instruction in Taiwan. *Journal of Research in Education Sciences, 65*(1), 93–114.] https://doi.org/10.6209/JORIES.202003_65(1).0004

國家教育研究院（無日期）：〈臺灣學生成就長期追蹤評量計畫網站〉。https://tasal.naer.edu.tw/

[National Academy for Educational Research. (n.d.). *Taiwan Assessment of Student Achievement: Longitudinal Study website*. https://tasal.naer.edu.tw/]

曾芬蘭、林奕宏、邱佳民（2017）：〈監控評分者效果的 Yes/No Angoff 標準設定法之效度檢核：以國中教育會考數學科為例〉。《測驗學刊》，*64*（4），403–432。[Tseng, F.-L., Lin, Y.-H., & Chiou, J.-M. (2017). Validation of rater-effects-monitored yes/no Angoff standard-setting method: Using the Taiwan comprehensive assessment program for junior high school students' math exam as an example. *Psychological Testing, 64*(4), 403–432.]

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。[Cheng, C.-M., & Wang, H.-P. (2012a). Assessing the standards set by TASA and its standard-setting procedures. *Educational Review, 39*, 77–118.] https://doi.org/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.]

黃馨瑩、謝名娟、謝進昌（2013）：〈臺灣學生學習成就評量英語科標準設定之效度評估研究〉。《教育與心理研究》，*36*（2），87–112。[Huang, H.-Y., Hsieh, M.-C., & Hsieh, J.-C. (2013). Validating the yes/no Angoff standard setting procedure on Taiwan Assessment of Student Achievement English test. *Journal of Education & Psychology, 36*(2), 87–112.]

謝名娟（2013）：〈以多層面 Rasch 分析的角度來評估標準設定之變異性〉。《教育心理學報》，*44*（4），793–811。[Hsieh, M.-C. (2013). Evaluating the variability in standard setting using many faceted Rasch model. *Bulletin of Educational Psychology, 44*(4), 793–811.]

謝進昌、謝名娟、林世華、林陳涌、陳清溪、謝佩蓉（2011）：〈大型資料庫國小四年級自然科學習成就評量標準設定結果之效度評估〉。《教育科學研究期刊》，*56*（1），1–32。[Hsieh, J.-C., Hsieh, M.-C., Lin, S.-H., Lin, C.-Y., Chen, C.-H., & Hsieh, P.-J. (2011). Validation of the standard setting procedure for a large scale 4th grade science assessment. *Journal of Research in Educational Sciences, 56*(1), 1–32.]

謝進昌（計畫主持人）（2021a）：《第四學習階段英語文素養長期追蹤》（計畫編號：NAER-2019-041-A-1-1-E1-07）。國家教育研究院年度研究成果報告，國家教育研究院。https://www.naer.edu.tw/bin/home.php [Hsieh, J.-C. (Principal Investigator). (2021a). *Longitudinal study of Taiwan junior high school students' English core competence* (Report No. NAER-2019-041-A-1-1-E1-07) (Grant). National Academy for Educational Research. https://www.naer.edu.tw/bin/home.php]

謝進昌（2021b）：《「混合專家與學生實徵表現導向」大型教育評量標準設定之效度評估研究》（已投稿），國家教育研究院測驗及評量研究中心。[Hsieh, J.-C. (2021b). *Assessing the validity of standard setting of large-scale assessment for English as a foreign language students with a hybrid*

of expert and empirical performance model (Manuscript submitted for publication). Research Center for Testing and Assessment, National Academy for Educational Research.]

Adams, R. J., Wilson, M. R., & Wang, W. L. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement, 21*(1), 1–24. https://doi.org/10.1177/0146621697211001

American Educational Research Association, American Psychological Association, & National Council on Measurement in Education. (2014). *Standards for educational and psychological testing*. American Educational Research Association.

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

Ardasheva, Y., Wang, Z., Adesope, O. O., & Valentine, J. C. (2017). Exploring effectiveness and moderators of language learning strategy instruction on second language and self-regulated learning outcomes. *Review of Educational Research, 87*(3), 544–582. https://doi.org/10.3102/0034654316689135

Baker, L., & Brown, A. L. (1984). *Metacognitive skills and reading*. In P. D. Pearson (Ed.), *Handbook of reading research* (pp. 353–394). Longman.

Barjesteh, H., Mukundan, J., & Vaseghi, R. (2014). Synthesis of language learning strategies: Current issues, problems and claims made in learner strategy research. *Advances in Language & Literary Studies, 5*(6), 68–74. https://doi.org/10.7575/aiac.alls.v.5n.6p.68

Beaton, A. E., & Allen, N. L. (1992). Interpretation scales through scale anchoring. *Journal of Educational Statistics, 17*(2), 191–201. https://doi.org/10.3102/10769986017002191

Beuk, C. H. (1984). A method for reaching a compromise between absolute and relative standards in examinations. *Journal of Educational Measurement, 21*(2), 147–152. https://doi.org/10.1111/j.1745-3984.1984.tb00226.x

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*(4), 431–444. https://doi.org/10.1177/014662168200600405

Bourque, M. L. (2009). *A history of NAEP achievement levels: Issues, implementation, and impact 1989-2009* (ED509389). ERIC. https://files.eric.ed.gov/fulltext/ED509389.pdf

Cizek, G. J., Bunch, M. B., & Koons, H. (2004). Setting performance standards: Contemporary methods. *Educational Measurement: Issues and Practice, 23*(4), 31–50. https://doi.org/10.1111/j.1745-3992.2004.tb00166.x

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

Cohen, A. S., Kane, M. T., & Crooks, T. J. (1999). A generalized examinee-centered method for setting standards on achievement tests. *Applied Measurement in Education, 12*(4), 343–366. https://doi.org/10.1207/S15324818AME1204_2

Cohen, J. (1988). *Statistical power analysis for the behavioral sciences* (2nd ed.). Lawrence Erlbaum Associates Publishers.

Efron, B. (1981). Nonparametric estimates of standard error: The jackknife, the bootstrapping and other methods. *Biometrika, 68*(3), 589–599. https://doi.org/10.1093/biomet/68.3.589

Gambrell, L. B., & Bales, R. J. (1986). Mental imagery and the comprehension-monitoring performance of fourth- and fifth- grade poor readers. *Reading Research Quarterly, 21*(4), 454–464. https://doi.org/10.2307/747616

Griffiths, C. (2007). Language learning strategies: Students' and teachers' perceptions. *ELT Journal, 61*(2), 91–99. https://doi.org/10.1093/elt/ccm001

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

Hambleton, R. K., & Plake, B. S. (1995). Using an extended Angoff procedure to set standards on complex performance assessments. *Applied Measurement in Education, 8*(1), 41–55. https://doi.org/10.1207/s15324818ame0801_4

Hu, L., & Bentler, P. M. (1999). Cutoff criteria for fit indexes in covariance structure analysis: Conventional criteria versus new alternatives. *Structural Equation Modeling, 6*(1), 1–55. https://doi.org/10.1080/10705519909540118

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

Jaeger, R. M. (1991). Selection of judges for standard-setting. *Educational Measurement: Issues and Practice, 10*(2), 3–14. https://doi.org/10.1111/j.1745-3992.1991.tb00185.x

Jeon, E. H., & Yamashita, J. (2014). L2 reading comprehension and its correlates: A meta-analysis. *Language Learning, 64*(1), 160–212. https://doi.org/10.1111/lang.12034

Kaftandjieva, F. (2010). *Methods for setting cut scores in criterion-referenced achievement tests: A comparative analysis of six recent methods with an application to tests of reading in EFL*. EALTA. https://www.ealta.eu.org/documents/resources/FK_second_doctorate.pdf

Kane, M. T. (1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*(3), 425–461. https://doi.org/10.2307/1170678

Kane, M. T. (2001). So much remains the same: Conception and status of validation in setting standards. In G. J. Cizek (Ed.), *Standard setting: Concepts, methods, and perspectives* (pp. 53–88). Erlbaum.

Kelly, D. L. (1999). *Interpreting the Third International Mathematics and Science Study (TIMSS) achievement scales using scale anchoring* [Unpublished doctoral dissertation]. Boston College.

Kolen, M. J., & Brennan, R. L. (2004). *Test equating, scaling, and linking: Methods and practices* (2nd ed.). Springer Publishing Company.

Linacre, J. M. (2005). *A user's guide to Winsteps/Ministeps: Rasch-Model programs*. MESA Press.

Martin, M. O., Mullis, I. V. S., Arora, A., & Preuschoff, C. (2014). Context questionnaire scales in TIMSS and PIRLS 2011. In L. Rutkowski, M. von Davier, & D. Rutkowski (Eds.), *Handbook of international large-scale assessment* (pp. 299–316). Chapman & Hall/CRC.

Masters, G. N., & Wright, B. D. (1997). The partial credit model. In W. J. van der Linden & R. K. Hambleton (Eds.), *Handbook of modern item response theory* (pp. 101–121). Springer Publishing Company.

McDonald, R. P., & Ho, M.-H. R. (2002). Principles and practice in reporting structural equation analyses. *Psychological Methods, 7*(1), 64–82. https://doi.org/10.1037/1082-989X.7.1.64

Mullis, I. V. S., Cotter, K. E., Centurino, V. A. S., Fishbein, B. G., & Liu, J. (2016). Using scale anchoring to interpret the TIMSS 2015 achievement scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in TIMSS 2015* (pp. 14.1–14.47). Boston College, TIMSS & PIRLS International Study Center.

Mullis, I. V. S., & Prendergast, C. O. (2017). Using scale anchoring to interpret the PIRLS and ePIRLS 2016 achievement scales. In M. O. Martin, I. V. S. Mullis, & M. Hooper (Eds.), *Methods and procedures in PIRLS 2016* (pp. 13.1–13.23). Boston College, TIMSS & PIRLS International Study Center.

Nassif, P. M. (1978, March). *Standard setting for criterion referenced teacher licensing tests* [Paper presentation]. National Council on Measurement in Education Annual Meeting, Toronto, Canda. https://www.ncme.org

Organisation for Economic Co-operation and Development. (2019). *PISA 2018 assessment and analytical framework.* https://doi.org/10.1787/b25efab8-en

Organisation for Economic Co-operation and Development. (2020). *PISA 2018 technical report*. https://www.oecd.org/pisa/data/pisa2018technicalreport/

Padron, N.Y., & Waxman, H. C. (1988). The effect of ESL students' perception of their cognitive strategies on reading achievement. *TESOL Quarterly, 22*(1), 146–150. https://doi.org/10.2307/3587068

Plonsky, L. (2011). The effectiveness of second language strategy instruction: A meta-analysis. *Language Learning, 61*(4), 993–1038. https://doi.org/10.1111/j.1467-9922.2011.00663.x

Robitzsch, A., Kiefer, T., & Wu, M. (2020). *TAM: Test analysis modules. R package version 3.5–19.* The Comprehensive R Archive Network. https://CRAN.R-project.org/package=TAM

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. https://doi.org/10.1080/08957340903221659

Warm, T. A. (1989). Weighted likelihood estimation of ability in item response theory. *Psychometrika, 54*, 427–450. https://doi.org/10.1007/BF02294627

Yao, L., & Schwarz, R. D. (2006). A multidimensional partial credit model with associated item and test statistics: An application to mixed-format test. *Applied Psychological Measurement, 30*(6), 469–492. https://doi.org/10.1177/0146621605284537

Zieky, M. J., & Livingston, S. A. (1977). *Manual for setting standards on the Basic Skills Assessment tests*. Educational Testing Service.

收稿日期：2020 年 10 月 07 日
一稿修訂日期：2020 年 12 月 13 日
二稿修訂日期：2020 年 12 月 22 日
三稿修訂日期：2020 年 12 月 28 日
接受刊登日期：2020 年 12 月 28 日

附件　本研究英語文理解策略使用自陳式題項

向度	題號	元素	題項內容
記憶策略	M1	圖像化	在學英文單字時，我會試圖把單字的發音和單字代表的圖像連結起來（例如，念 refrigerator（冰箱）時，腦袋會聯想到冰箱的樣子）
	M2	聲韻	我會用相似的發音（例如，old, cold, hold 都有 o 的長母音）來記憶新的英文單字
	M3	脈絡化	我會將剛剛新學到的英文單字，轉換為自己熟悉或想像的場景，以幫助記憶
	M4	搭配詞／同義詞	在背單字時，我會將意思有關聯的英文單字都放在一起以幫助記憶（例如，good 與 excellent 都表示「好」的意思）
	M5	字根／首／尾	當碰到不熟悉的英文單字時，我會用字首、字根、字尾的概念，把英文單字拆成幾個部分再背起來（例如，player 是 play 與 er 組成）
	M6	音節	當碰到不熟悉的英文單字時，我會根據音節將英文單字拆成幾個部分再背起來（例如，sunny 會分成 sun-ny 兩個音節）
認知策略	C1	快速瀏覽	讀英文時，我會先大致讀過整篇文章，了解文章的大意後再仔細閱讀
	C2	重點標示／畫底線	上英文課時，我會將上課聽到或讀到的內容重點標示出來（例如，畫底線）
	C3	摘取大意	我會把英文文章中重要的內容或訊息簡短地記錄下來
	C4	文法規則	讀英文時，我會試著找出英文片語或句型的用法與規則來幫助我理解文章大意
	C5	語調／氣	在聽別人說英語時，我會利用不同的方式（例如，語調、或停頓語氣），來了解對方所說的內容
	C6	文本架構	讀英文時，我會先分析文章內容的架構（例如，文章長度、主要人物）後，再仔細閱讀
推論連結策略	I1	上下文脈絡	我會從上下文或整段內容來理解文章在說甚麼
	I2	圖表／標題	我會從英文文章中插圖、表格或標題來推測文章大意
	I3	上下文脈絡	用英語交談或閱讀時，我會利用上下文已經知道的字（或其它線索），來推測新字或詞的意思
	I4	大意理解	讀英文時，我會先了解大意而不會碰到每個不懂的單字都去查字典
	I5	關鍵字詞	進行英語交談時，我會根據聽到的一些關鍵字詞來了解大意
	I6	肢體動作／表情	進行英語交談時，我會注意對方的臉部表情或肢體語言（例如，手勢）來幫助了解他想表達的訊息
	I7	背景知識／經驗	在聽或讀英文時，如果碰到不懂的內容時，我會試著從自己過去的經驗與知識來推測
	I8	概念關係連結	我會來回閱讀文章內容，試著找出文章裡不同概念之間的關係
理解監控策略	S1	監控	我會留意自己英語文學習的進步情況
	S2	調整	我會記下自己常犯的英語文錯誤並改進，以幫助學習
	S3	檢驗	我會試著用自己理解的方式，重新檢查所聽或讀到的英語文內容，以幫助我更加了解
	S4	檢驗	如果我找到的資訊對於同一件事情有不同的說法（例如，有人說吃蛋會導致高血壓，有些人則說兩者沒有關係），我會再次確認這些資訊的來源是否正確
	S5	檢驗	當碰到英文內容與我知道的不一致時，我會再次確認內容是否正確
	S6	監控	學習英語文時，我會一邊讀一邊確認自己哪裡不懂
	S7	監控	在聽和讀英語文時，當我發現自己不專心時，我會再集中注意力來聽和閱讀內容
	S8	調整	當碰到英語文相關問題時，我會發展出自己理解或解決問題的方法
	S9	調整	學習英語文遇到不懂的地方時，我會想辦法去解決
	S10	評估	我會小心地評估與選擇我找到的相關英文素材（例如，網路文章、雜誌），才進行閱讀

註：本問卷題項發展源自於研究案〈第四學習階段英語文素養長期追蹤〉（編號：NAER-107-12-B-1-07-06-1-08），謝進昌，2021a，國家研究院（https://www.naer.edu.tw/bin/home.php）。

Bulletin of Educational Psychology, 2021, 53(2), 307–334
National Taiwan Normal University, Taipei, Taiwan, R. O. C.

Extended Angoff Method in Setting Standards for Self-Report Measures With Supplementary Performance-Level Descriptors

Jin-Chang Hsieh
Research Center for Testing and Assessment,
National Academy for Educational Research

One vision of the 12-Year Basic Education Curriculum in Taiwan is to promote the comprehensive learning and development of all students. To ensure the quality of this curriculum reform, the Ministry of Education funded a long-term project, the Taiwan Assessment of Student Achievement: Longitudinal Study (TASAL), to evaluate the impact of the curriculum on student performance. The TASAL is a large-scale standards-based assessment, and standard setting is one of its main task. A comprehensive literature review indicated that most empirical studies related to standard setting have focused on cognitive domains and few study undertake expert-oriented standard-setting processes in affective domains because of some practical limitations. The present study suggests a new approach, employing supplementary performance-level descriptors (S-PLDs) in an extended Angoff method in setting standards for self-report measures. The purpose of this study was to uncover evidence of the procedural, internal, and external validity of implementing an extended Angoff method procedure with S-PLDs in standard setting for English comprehension strategy use among seventh grade students in Taiwan.

PLDs are designed to outline the knowledge, skills, and practices that indicate the level of student performance in a target domain. In the present study, the use of comprehension strategies for learning English as a foreign language was examined. S-PLDs provide comparable but unique functions within the standard-setting process. S-PLDs offer supplementary material to subject matter experts to facilitate the formation of profiles of student performance in target domains, especially when ambiguities in conventional PLDs may prevent expert consensus during the standard-setting process.

In this study, stratified two-stage cluster sampling was adopted to select representative seventh graders in Taiwan during the 2018–2019 academic year. After sampling, 7,246 students had been selected ; only 2,732 students, 1,417 boys and 1,315 girls, received an English comprehension strategy use questionnaire and English proficiency test. Student performance on both measurement instruments was the basis for writing PLDs and S-PLDs. The scale measuring English comprehension strategy use was a 4-point discrete visual analogue scale self-report measure developed through standardized procedures and comprises four dimensions: memorization (6 items), cognition (6 items), inference (8 items), and comprehension monitoring (10 items) strategies. The results of four-dimensional confirmatory factor analysis indicated a favorable model–data fit, except for the chi-square value, which was affected by the large sample size . Moreover, the English proficiency test used was a cognitive measure assessing students' listening and reading comprehension abilities through the use of multiple-choice and constructed-response items. A total of 182 items were developed through a standardized procedure and divided into 13 blocks to assemble 26 test booklets. Each booklet, containing 28 items, was randomly delivered to a participating student; each student completed only one booklet. After data cleansing and item calibration with a multidimensional random coefficient multinomial logit model and the Test Analysis M odules (Robitzsch et al., 2020), the information-weighted fit mean-square indices for all test items ranged from

0.79 to 1.37, meeting the criterion proposed by Linacre (2005).

An expert-oriented standard-setting meeting was hosted on May 20, 2020, after advanced materials, such as agenda, instruction of standard-setting method, had been sent to all experts. Eight experts from across Taiwan were invited to join the meeting, and they all had experience involving standard-setting meetings for student performance on English proficiency tests. The average number of years of teaching experience for these experts was 18.75, and seven had experience in teaching low achievers. Overall, the experts had sufficient prerequisite knowledge and experience with standard-setting processes. On the day of the standard-setting meeting, a series of events, including orientation, training and practice, and three rounds of extended Angoff standard-setting methods with different types of feedback provided between rounds, were undertaken. Feedback questionnaires were developed , and discussions among the experts between the rounds were recorded and analyzed as evidence of procedural and internal validity.

Most of the subject matter experts were satisfied with the events during the standard-setting process and agreed that they could set satisfactory cutoff scores for future usage. From the results of feedback questionnaires completed between rounds, the experts nearly unanimously agreed that the materials received in advance; the introductions to PLDs, S-PLDs, and the extended Angoff method; and previous experience in setting standards for English proficiency were beneficial in judging items during the process. Additionally, the experts agreed that the S-PLDs played a key role in facilitating the formation of outlines for student performance in comprehension strategy use across different levels. All of these results indicate procedural validity.

For evidence of internal validity, classification error (the ratio of the standard error of the passing score to the measurement error), was computed to indicate the consistency of the item ratings between and within the experts during the three-round process. Between experts and across rounds, the classification error ranged from 0.08 to 0.36 for memorization strategies, 0.14 to 0.49 for cognition strategies , 0.19 to 0.61 for inference strategies, and 0.24 to 0.72 for comprehension monitoring strategies. These results indicate that the cognitive levels for the four dimensions affect the consistency of item rating. Strategies with more abstract item content tended to have higher classification error. Furthermore, the lowest classification error values occurred in the second round for memorization and inference strategies and in the third round for cognition and comprehension monitoring strategies. All low values for each dimension were beneath the cutoff of 0.33 proposed by Kaftandjieva (2010), except for the value of 0.37 for comprehension monitoring strategystrategies. Regarding the rating consistency within experts between rounds, the results showed no extreme classification error, and most of the values were beneath 0.33, with the exceptions of 0.35 for cognition strategies and 0.37, 0.42, and 0.61 for comprehension monitoring strategies. Therefore, most experts exhibited rating consistency between the rounds. Additionally, the results of a content analysis of the item rating discussions indicated that three reference sources might affect experts' judgments regarding the items: (1) students' actual performance, (2) PLDs and S-PLDs, and (3) experts' personal expectations. For example, one expert might give a lower score because his students tend to exhibit poor performance on a particular item dependent on their teaching experience, whereas another expert might give a higher score because of their personal expectations.

To examine external validity, student performance on English proficiency tests was adopted as an external criterion . With two cutoff scores used to divide students into basic, proficient, and advanced users in each dimension, a medium effect size was obtained for memorization strategies, and large effect sizes were obtained for cognition, inference, and comprehension monitoring strategies. Furthermore, to compare the final cutoff scores obtained through the study method with existing methods , the study adopted the concept from TIMSS and PIRLS for setting standards for affective domains (Martin et al., 2014, p. 308). The classification accuracy indices, which indicate the proportions of students classified identically, were 90.25%, 81.20%, 82.52%, and 87.56% for the four dimensions. To sum up, the present study obtained satisfactory evidence of the procedural, internal, and external validity of using an extended Angoff procedure for setting standards for self-report measures with S-PLDs; additional suggestions are presented herein.

Keywords: extended Angoff method, self-report measure, supplementary performance-level descriptors, standard setting, Taiwan Assessment of Student Achievement: Longitudinal Study