華語文寫作測驗垂直等化研究

藍珮君[1]、王從安[2]、彭淑惠[3]、陳柏熹[4]、曾文璇[5]
國家華語測驗推動工作委員會[1*235]
國立臺灣師範大學教育心理與輔導學系、華語文與科技研究中心[4]
martinalan@sc-top.org.tw[*]

摘要

華語文寫作測驗是專為母語非華語人士研發之語言測驗。華語文寫作測驗為一分級測驗，將華語文寫作能力分成三等六級，三等分別為入門基礎級、進階高階級及流利精通級，而每一等又可再依據測驗成績細分為兩級，分別為入門級、基礎級、進階級、高階級、流利級、精通級，共六級。

華語文寫作測驗在每個測驗等級依據語言能力的發展以及題型特性，分別制定不同的評分原則以評量考生的作答反應。因此，跨等級的測驗成績無法直接進行比較，進而在了解跨不同測驗等級應試者表現有無進步或進步幅度方面有所限制。

為了解決此一問題，華語文寫作測驗的研發單位華測會，在辦理測驗改版預試的同時，進行寫作測驗的垂直等化研究，希冀了解入門基礎級、進階高階級與流利精通級不同級分的對應關係，俾利跨等級的測驗成績可相互比較。

本研究採用定錨題不等組設計（non-equivalent groups with anchor test design，簡稱 NEAT），招募 135 名研究參與者，依照參與者自選組別分為兩組，分別作答兩份測驗，兩份測驗中有一道內容相同的定錨題作為資料串連之用。此外，另搭配過去兩次正式考試共 72 名應試者之作答文本，以完成三個測驗等級的資料串連。彙整所有作答反應文本完成評分後，使用 ConQuest 軟體進行潛在迴歸分析（latent regression analysis），估計考生能力值與試題難度參數。

本研究結果發現：1. 不同組別應試者的整體能力存在差異；2. 三個測驗等級試題難度適合用來區分不同程度的應試者；3. 建立不同測驗等級級分之間的對應關係。上述結果有助於教學實務層面了解跨不同測驗等級的學習者能力發展情形。對於學習者來說，可以了解自身華語文寫作能力的進步情形，判斷自己的能力有無成長；對於教學者來說，可以藉此評估教學對象未來應針對目前等級持續加強，或是可朝向下一個等級學習。

關鍵字：華語文能力測驗、寫作測驗、垂直等化、潛在迴歸分析

一、前言

華語文寫作測驗是專為母語非華語學習者所設計，由國家華語測驗推動工作委員會（以下簡稱華測會）專責研發。本測驗以溝通任務為導向，在等級規劃方面，將華語文寫作能力分成三等六級，三等分別為入門基礎級、進階高階級與流利精通級，而每一等又可再依據測驗成績細分為兩級，分別為入門級、基礎級、進階級、高階級、流利級、精通級，共六級。測驗方式為採用電腦打字輸入的寫作方式。

華語文寫作測驗在每個測驗等級依據語言能力的發展、可處理的主題範圍、溝通任務的複雜度，制定相對應的題型，並據此分別擬定不同的評分原則以評量應試者的作答反應，因此，跨不同測驗等級的級分無法直接進行比較。例如：在入門基礎級評分原則得到最高級分 5 級分的應試者，其表現相當於進階高階級的哪一級分；或是在進階高階級評分原則得到最高級分 5 級分的應試者，若報考流利精通級可能獲得什麼樣的成績。進而在了解跨不同測驗等級的應試者表現有無進步，抑或是進步程度有多少有所限制，甚為可惜。

華測會為了探討此一問題，在 2022 年辦理華語文寫作測驗改版預試的同時，進行了寫作測驗的垂直等化研究，欲探討以下三個問題：

1. 垂直等化中不同組別應試者能力是否有差異。
2. 入門基礎級、進階高階級與流利精通級試題難度是否有差異。
3. 了解並建立入門基礎級、進階高階級與流利精通級不同級分的對應關係。

二、文獻探討

2.1 華語文寫作測驗改版介紹

因應國家教育研究院制定之「臺灣華語文能力基準」於 2020 年發布（林慶隆等，2020），以及歐洲共同語文參考架構（Common European Framework of Reference for Languages，簡稱 CEFR）2018 年在部份等級新增能力指標（Council of Europe, 2018），華語文寫作測驗自 2021 至 2022 年進行測驗改版，調整部分測驗題型。評分原則方面，各測驗等級題型仍採用分析式評分，但縮減了評分向度，調整為任務完成度、文意連貫性以及形式語言適切性三大向度。以下針對能力描述以及測驗題型進行說明。

2.1.1 能力描述

改版後的華語文寫作測驗能力指標如表 1 所示，調整的等級為入門級、基礎級、進階級與高階級，流利級與精通級因未更動故在此省略。與先前的能力指標相較，此次改版強調各個通過等級在不同溝通任務可達到的語言表現。以入門級為例，前一版能力描述為「能寫出簡單、不連貫的短語和句子」，新版改為「能以短語和簡單、不連貫的句子描述與日常生活相關的人物、地點、物品及活動」以及「能以短語和簡單、不連貫的句子書寫與日常生活相關、簡短且非常簡單的訊息，提供資訊或詢問」，更加具體描述達到入門級水準的學習者寫作表現。

表 1　改版華語文寫作測驗能力描述

測驗等級	通過等級	能力描述
入門基礎級	入門級	● 能以短語和簡單、不連貫的句子描述與日常生活相關的人物、地點、物品及活動。 ● 能以短語和簡單、不連貫的句子書寫與日常生活相關、簡短且非常簡單的訊息，提供資訊或詢問。
	基礎級	● 能以簡單連接詞銜接數個句子，如：而且、可是、因為……。 ● 能書寫文意連貫的數個句子簡單地描述與日常生活相關的事件、活動和個人經驗。 ● 能書寫文意連貫的較短的信件來交換訊息、回應日常生活相關的問題或表達立即的需求。
進階高階級	進階級	● 能書寫淺白、連貫的文本。 ● 能書寫信件詢問資訊、提出要求、給予確認，並提供有限的相關細節，或較詳細的描述經驗、情感、事件。 ● 對一般性社會議題，能書寫簡短的文章，敘述事實資訊及行為的原因，能簡單說明自己的立場與觀點。
	高階級	● 能書寫清楚、詳細的文本。 ● 能以適切的格式與語體書寫信件，傳達不同程度的情緒，強調事件和經驗對個人的意義，並對通信者傳達的消息或觀點給予評論。 ● 對一般性社會議題，能以適切的例證闡述個人支持或反對某特定觀點的理由。

2.1.2 測驗題型

測驗題型方面，配合能力描述微調，原入門基礎級第一部分改為「圖片描述」題型，考生需依據題目提供的兩張圖片內容，寫出一篇 100 字左右的短文，第二部分「書信寫作」題不變；原進階高階級第一部分仍為「書信寫作」題，題型未變動，但題目主題範圍除了私人書信表達情感與事件以外，更拓展至正式、實用書信主題，如工作、客訴或諮詢信件等，且所設計的寫作任務能讓進階級與高階級能力考生均能有所發揮，第二部分「觀點論述」題不變。由於流利精通級測驗能力描述未調整，因此測驗題型也維持現行「摘要寫作」、「觀點論述」兩題型。

下表 2 為預試時華語文寫作測驗題型說明，入門基礎級兩個題型作答時間均為 20 分鐘，字數要求皆為 100 字左右，考量此等級應試者閱讀能力較弱，注意事項的說明提供英、法、德、西、日、韓、越、泰、印尼九國語言的翻譯輔助供其選擇；進階高階級兩個題型作答時間各為 60 分鐘，也就是 1 小時，字數要求為 600 字左右，在此僅用於說明預試時所要求字數與時間，最終改版要求字數，請以華測會官網公告為準。圖 1 與圖 2 為兩個測驗等級的題型例題。

表 2　華語文寫作測驗題型（預試）

測驗等級	部分	寫作任務	字數要求	作答時間
入門基礎級	1	圖片描述	100 字左右	20 分鐘
	2	書信寫作	100 字左右	20 分鐘
進階高階級	1	書信寫作	600 字左右	60 分鐘
	2	觀點論述	600 字左右	60 分鐘

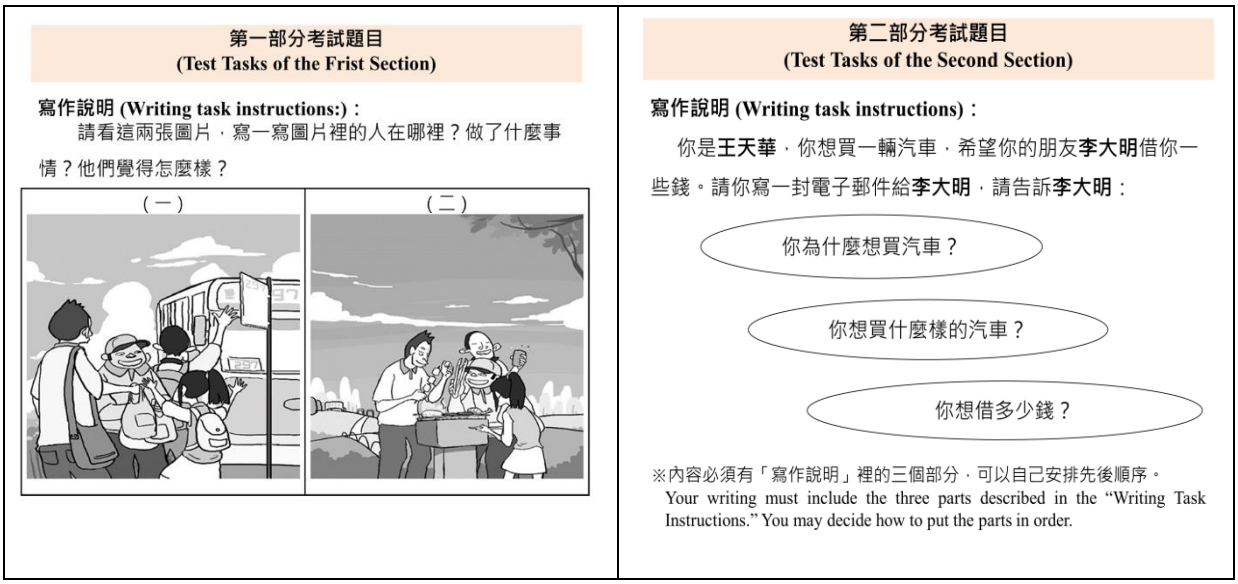

圖 1　入門基礎級寫作測驗例題

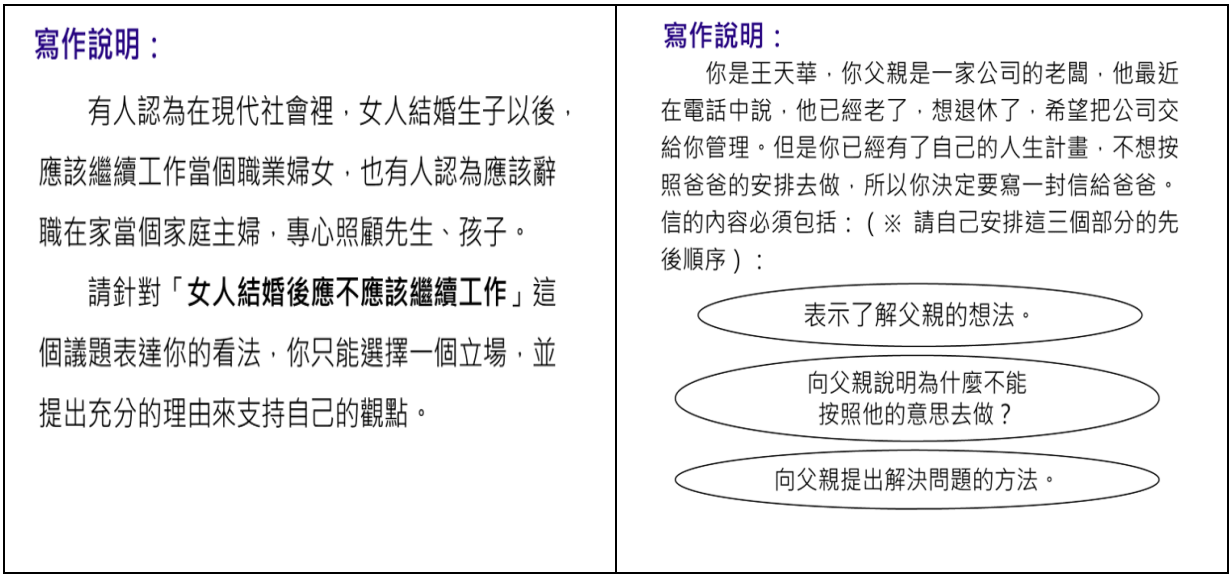

圖 2　進階高階級寫作測驗例題

2.2 測驗等化

2.2.1 等化種類

等化（equating）是一種統計程序，用來調整測驗形式的分數，讓不同形式的測驗

分數可以互換使用（interchangeably），以便比較不同形式測驗所獲得的分數（Kolen & Brennan, 2004）。Hambleton 與 Swaminathan（1985）指出測驗等化可分為水平等化（horizontal equating）與垂直等化（vertical equating）兩種。前者是指對兩個以上測量相同特質、相同能力且難度相近的測驗，將其原始分數轉換至同一量尺的過程，常被應用在許多大型測驗，如：托福、GRE，與基本學力測驗等考試；後者是指對兩個以上測量相同特質、相同能力但難度不一的測驗，將其原始分數轉換至同一量尺的過程（引自張鈺卿，2007）。

垂直等化可以將測量特質或能力範圍較廣的測驗結果進行相互比較，當測驗測量到同一特質且包括不同程度，研究者又希望這些不同程度的測驗能使用相同的計分量尺時，便適合使用此一方式。本研究目的為探討入門基礎級、進階高階級與流利精通級寫作測驗不同級分的對應關係，也就是要將測量特質同為華語文寫作能力但難度不同之測驗轉換至同一量尺，故屬於垂直等化。

2.2.2 等化設計

測驗等化設計是指收集等化資料的方法。一般常見的等化設計包括單組設計（single group design）、平衡對抗隨機組設計（counterbalanced equivalent groups design）、等群組設計（equivalent group design）、平衡不完全區塊設計（balanced incomplete block design，簡稱 BIB）、試題預先等化設計（item pre-equating design），以及定錨題不等組設計（non-equivalent groups with anchor test design，簡稱 NEAT）等（王寶墉，1995；余民寧，2009；Kolen & Brennan, 2004）。

定錨題不等組設計為兩組不同考生作答兩份題本，題本之中放置相同試題，藉由相同試題將兩份題本的其他試題進行連結。華語文寫作測驗題數雖然不多，但除了入門基礎級測驗以外，其他測驗等級單一題型的作答時間需要 50 或 60 分鐘以上，為避免研究參與者因測驗時間過久產生疲勞感進而影響作寫作表現，將採用定錨題不等組設計進行等化研究。

三、研究方法

3.1 分析資料來源

本研究分析資料有兩大來源，一是 2022 年 4 月參加寫作測驗改版預試的考生，甲群有 56 人，乙群有 79 人；另一是過去正式考試的考生，丙群為進階高階級 30 人，丁群為流利精通級 42 人。合計有 207 名應試者，皆為母語非華語的外籍人士。

3.2 研究設計

等化設計上採用定錨題不等組設計的概念，甲群應試者回答兩題入門基礎級試題加上一題進階高階級「書信寫作」題；乙群則回答兩題進階高階級試題再加上一題流利精通級「摘要寫作」題，其中，進階高階級測驗題目與丙群相同。透過將進階高階級「書信寫作」題作為定錨題的方式，連結甲群與乙群應試者的作答反應。

此外，乙群應試者回答的「摘要寫作」題挑選與丁群應試者使用的流利精通級正式考試相同題目，作為另一道定錨題，串聯乙群與丁群應試者作答反應。如此便可藉由兩個不同測驗等級的定錨題完成華語文寫作測驗三個等級的作答反應連結。

3.3 分析方法

本研究採用試題反應理論（item response theory）分析軟體 ConQuest 進行單向度潛在迴歸分析（unidimensional latent regression analysis）。主要考量為作答反應資料來自三個測驗等級共四個群組，研究團隊認為不同群組之整體能力應有所差別，不宜以視為單

一群體進行垂直等化分析。而潛在迴歸分析的特點為可將反應資料依照作答者潛在特質或能力分成二至數個群組，比較不同群組整體能力值的差異（Wu, Adams & Wilson, 1998），故採用此法，將甲、乙、丙、丁四群體標記為自變項進行潛在迴歸分析。

四、研究結果與討論

4.1 不同群組能力值差異

研究團隊以四個群組進行單向度潛在迴歸分析，分析結果顯示甲群應試者的平均能力值為-2.485 logits，而乙群整體能力值較甲群高 3.418 logits，為 0.933 logits，丙群與丁群分別較甲群能力值高出 2.644 logits 和 5.089 logits，整體能力值為 0.159 logits 和 2.604 logits。

上述結果顯示不同群組考生能力值有所不同，符合研究團隊預期。甲群應試者作答試題為等級最低的入門基礎級與一題進階高階級書信寫作題，整體試題難度最低。乙群作答試題除進階高階級外，尚包含一題流利精通級摘要寫作題，且在進階高階級測驗原始級分表現平均達 3.9 級分；丙群僅作答進階高階級測驗且原始平均級分表現為 3.3 級分，兩者相比乙群整體表現略優於丙群。最後，丁群作答測驗等級最高的流利精通級試題，整體能力表現為四組中最高。

4.2 不同測驗等級試題難度差異

表 3 為垂直等化分析得到的華語文寫作測驗六道試題的參數估計結果，題型名稱前 A、B、C 分別表示入門基礎級、進階高階級與流利精通級。由下表可知，最困難的試題為流利精通級試題，兩題難度相近，分別為 2.587 logits 與 2.577 logits；最簡單的試題為入門基礎級的書信寫作，難度為-2.842 logits。試題適配度方面，無論是訊息加權或未加權的均方差（MNSQ）均未符合一般對於建構反應題型介於 0.5 至 1.5 之間的標準，T 值亦大於一般可接受的範圍±3.0。對此研究團隊認為可能原因為採用潛在迴歸分析，得到的試題難度分布範圍較大，因而容易造成適配度不佳的情形。若從 MNSQ 的信賴區間數值來看，則大致上符合 0.5 至 1.5 之內的標準，試題適配度應為尚可接受。

進一步檢視測驗等級內的試題難度，可以發現同一測驗等級中難度參數均相差 0.564 logits 以內，差異不大；而不同測驗等級之間試題難度有較為明顯的差異，入門基礎級兩題試題難度低於-2 logits，進階高階級兩題難度落在-1 logits 左右，而流利精通級兩題難度均高於 2.5 logits。此外，可從試題區分信度（item separation reliability）檢視試題品質，此指標介於 0 至 1 之間，數值越高表示試題難度差異越明顯，更能良好地區分應試者能力差異（陳姿螢、洪素蘋、樂鍇祿璞崚岸，2019）。分析報表顯示數值達.997，顯示華語文寫作測驗試題能有效區辨不同程度的應試者。

表 3　華語文寫作測驗試題參數

題序	試題名稱	難度參數	標準誤	未加權適配指標			加權適配指標		
				MNSQ	信賴區間	T 值	MNSQ	信賴區間	T 值
1	A 圖片描述	-2.553	0.156	10.24	0.63~1.37	18.7	9.24	0.46~1.54	12.0
2	A 書信寫作	-2.842	0.154	9.69	0.63~1.37	18.0	10.09	0.41~1.59	11.7
3	B 書信寫作	-1.394	0.100	2.81	0.78~1.22	11.3	2.78	0.77~1.23	10.5

題序	試題名稱	難度參數	標準誤	未加權適配指標			加權適配指標		
				MNSQ	信賴區間	T 值	MNSQ	信賴區間	T 值
4	B 觀點論述	-0.830	0.119	1.37	0.73~1.27	2.5	1.42	0.72~1.28	2.7
5	C 摘要寫作	2.587	0.132	3.48	0.75~1.25	12.0	3.61	0.70~1.30	10.6
6	C 觀點論述	2.577	0.177	8.86	0.57~1.43	14.8	8.29	0.49~1.51	11.9

4.3 試題級分對應關係

下圖 3 為試題賽斯通閾值（item Thurstonian thresholds）對應圖，最左欄數字表示考生能力參數及試題閾值參數，而每一個 X 表示 0.9 名考生，右方數字代表試題與級分。舉例來說，3.5 表示第 3 題（B 書信寫作）的 5 級分，而其所對應到的數值接近 1，表示能力值為 1 logits 的應試者約有 50%的機會可以在此題得到 5 級分。由於此分析結果僅為單次收樣，且樣本數有限，故在試題級分的轉換上，研究團隊採用同一測驗等級兩道試題相同級分的能力值平均數來代表該測驗等級各級分的對應能力值，結果整理如表 4。

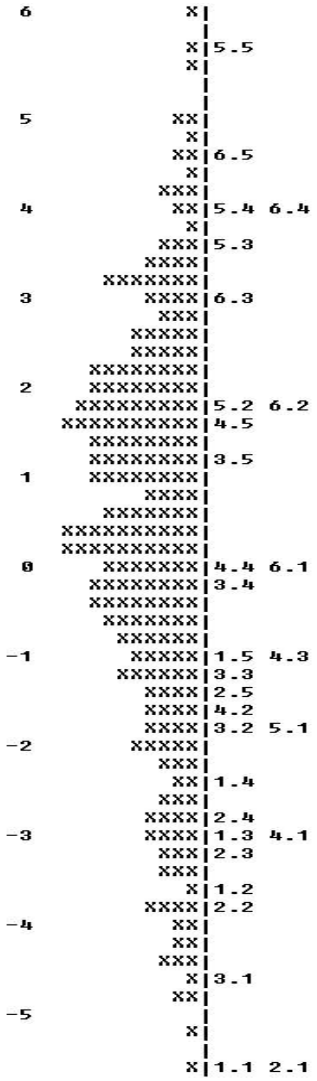

圖 3　試題塞斯通閾值對應圖

因不同等級試題主題範疇及要求的任務不同，不同等級題目可能存在試題差異功能，因此在推論上需特別謹慎。從表 4 看來，入門基礎級 2 級分（-3.75）接近於進階高階級 1 級分（-3.78），也就是說在入門基礎級得到 2 級分的應試者，在面對到進階高階級的試題時，有機會得到 1 級分。而在入門基礎級最高級分 5 級分的應試者，其能力值（-1.25）約接近於進階高階級 3 級分（-1.10），又接近流利精通級測驗 1 級分（-0.97），表示有機會在進階高階級得到 3 級分。研究團隊認為因 5 級分為入門基礎級測驗最高級分，表示此程度考生能力已超過入門基礎級試題可量測範圍，能力甚或可能具備進階高階級水準，故有此一結果。

此外，進階高階級 3 級分平均能力值（-1.10）約相當於流利精通級 1 級分（-0.97），顯示在進階高階級試題得到 3 級分的應試者，有機會在流利精通級試題得到 1 級分，但還未能達到 2 級分水準。進階高階級需要達到 5 級分（1.28）才約接近流利精通級 2 級分（1.69）。表示在進階高階級試題得到 5 級分的應試者，有機會在流利精通級試題得到 2 級分。同樣因 5 級分為進階高階級最高級分，寫作表現達此程度之應試者，能力水準已超過進階高階級試題可量測範圍，可能已具備流利精通級水準。

表 4　各測驗等級單題各級分平均能力值

入門基礎級 (級分)	平均能力值	進階高階級 (級分)	平均能力值	流利精通級 (級分)	平均能力值
A (1)	<-5.0	B (1)	-3.78	C (1)	-0.97
A (2)	-3.75	B (2)	-1.81	C (2)	1.69
A (3)	-3.21	B (3)	-1.10	C (3)	3.19
A (4)	-2.64	B (4)	-0.16	C (4)	3.91
A (5)	-1.25	B (5)	1.28	C (5)	5.11

註：A 等兩題型均無考生得到 0 級分，可能因此在估計 1.1 與 2.1 的 thresholds 時發散。

五、結論與建議

本研究針對華語文寫作測驗三個測驗等級共六道試題進行垂直等化，將入門基礎級、進階高階級、流利精通級試題放在同一量尺上比較，從中確認不同群組應試者整體能力有所差異，丁群（流利精通級正式考試）應試者能力值最高；其次為乙群考生（進階高階級加流利精通級預試）與丙群考生（進階高階級正式考試）；最後是甲群考生（入門基礎級加進階高階級預試）。也得知三個測驗等級試題難度有所差異，適合用來區分不同程度的應試者。

此外，更重要的是建立不同測驗等級級分之間的對應關係，此結果可應用在教學實務層面上，有助於了解跨不同測驗等級的學習者能力發展情形。對於學習者來說，可以了解自身華語文寫作能力的進步情形，判斷自己的能力有無成長。例如：參加入門基礎級測驗總分得到 9 分的學習者，表示其中一題得到 5 級分，另一題得到 4 級分，平均能力值約為-1.95 logits，表示其寫作表達能力已有長足的進步，可往下一個測驗等級，也就是進階高階級測驗邁進，有機會在此一等級得到相當於測驗總分 4 分左右的表現。

而對於教學者來說，可以藉此對應結果評估教學對象未來應針對目前等級持續加強，或是可朝向下一個等級學習。若學生在進階高階級總分得到 8 分，表示其還需要在

進階高階級範圍繼續加強；若測驗表現可達到滿級分 10 分，教學者便可規劃讓學習者跨到下一個等級進行更深入的學習。

六、參考文獻

1. 王寶墉（1995）。現代測驗理論。心理出版社，臺北市。
2. 余民寧（2009）。試題反應理論 IRT 及其應用。心理出版社，臺北市。
3. 林慶隆、吳鑑城、白明弘、李詩敏、吳欣儒、蔡岳璋、丁彥平、張玳維、余昱瑩、王冠孺、盧昱勳、陳沛璇（2020）。遣辭用「據」：臺灣華語文能力第一套標準，初版，國家教育研究院，新北市。
4. 張鈺卿（2007）。BIB 與 NEAT 設計在不同年度測驗連結效果之比較。國立臺中教育大學教育測驗統計研究所碩士論文（未出版）。
5. 陳姿螢、洪素蘋、樂鍇祿璞崚岸（2019）。文化回應教學量表編製，發表於 2019「多元族群教育與文化回應教學」國際學術研討會，台北，5 月 30 日至 6 月 1 日。http://210.240.179.19/2019MECRT/index.php/conference-manual-and-paper-download/
6. Council of Europe. (2018). Common European Framework of Reference for Languages: Learning, Teaching, Assessment – Companion Volume with New Descriptors. Strasbourg: Council of Europe Publishing. Retrieved Oct 5, from https://rm.coe.int/cefr-companion-volume-with-new-descriptors-2018/1680787989.
7. Kolen, Michael J. & Brennan, Robert J. (2004). Test Equating, Scaling and Linking: Methods and Practices. 2nd edition. Springer, New York.
8. Wu, M., Adams, R. J., & Wilson, M. R. (1998). Acer ConQuest. ACER Press, Melbourne.