華語文教學研究 16.2: 75-102, 2019

華語文寫作測驗信度與評分規準適切性研究
－以流利精通級摘要寫作題型為例*

陳柏熹
國立臺灣師範大學教育心理與輔導學系、學習科學跨國頂尖研究中心

彭淑惠
國家華語測驗推動工作委員會

藍珮君
國家華語測驗推動工作委員會

摘要

寫作測驗因題數少且須仰賴人工評分，使其信度研究更形重要，然而過往華語寫作測驗信度相關文獻甚為少見。緣此，本研究針對流利精通級華語文寫作測驗摘要寫作題型的信度與評分規準的適切性進行探討，運用的分析模式包括斯皮爾曼等級相關分析、類推性理論與多元迴歸分析。研究結果顯示：（1）在採取分析式評分方式的情況下，多數評分者所評定出的整體級分與最後成績呈中度或高度正相關，評分者間信度大致良好。（2）受試者的變異成分最高，可佐證其得分能反映寫作能力。（3）文本由 3 人評閱，有利兼顧評分品質與經濟效益。（4）在 11 個變項中，複雜句型、組織、詞彙語法、詞語運用、規範性為預測力較為顯著的變項。

關鍵詞：信度　流利精通級　華語文寫作測驗　評分規準　摘要寫作

* 本研究感謝由教育部補助之國立臺灣師範大學高等教育深耕計畫「學習科學跨國頂尖研究中心」團隊在研究分析上的專業協助，以及本刊兩位匿名審查委員之寶貴建議，特此致謝。

1. 前言

「華語文能力測驗」（Test of Chinese as a Foreign Language，簡稱 TOCFL）是檢測以中文為第二語言學習者的中文能力檢定考試，其證書可作為母語非華語人士的升學與就業資格的參考標準(Chang 2017)，因此測驗的信度(reliability)極為重要。就測驗類別而言，聽力測驗與閱讀測驗有固定答案，屬於客觀性測驗，而口語測驗與寫作測驗皆為實作評量，沒有標準答案，因此在評分上難以避免地具有較大的主觀性。其中，寫作測驗受試者完成一篇作文所需的時間較長，且人工評閱較為費時，受限於上述因素，通常每次只考 1 或 2 個題目。正因為測驗題數少，加上成績需仰賴評分者的判斷，使其評分結果能否真實反映寫作能力，經常受到外界的質疑（王文中、呂金燮、吳毓瑩、張郁雯、張淑慧 2008；陳柏熹 2011；王德蕙、李奕璇、曾芬蘭、宋曜廷 2013），然而測驗與教育學界對於華語文寫作測驗在評分規準(rubric)與評分方式的探討並不多見，信度方面的研究明顯不足。

「歐洲共同語文參考架構」（The Common European Framework of Reference for Languages: Learning, Teaching, Assessment，簡稱 CEFR）是國際間廣泛作為語文課程設計、教學方法、教材規劃與語言能力評量之參考框架，此架構將語文能力由 A1 至 C2 分為六級，C2 為能力最高等級。在寫作能力評量方面，將寫作活動與策略分為寫作表達(written production)、寫作互動(written interaction)與寫作文本形式(text-types)三大類。其中，第三大類包含記筆記(Note-Taking)與文本處理能力(Processing Text)，C1 之記筆記能力描述為「在聆聽自己感興趣的演說時，能詳盡且正確地記筆記，且貼近講者的原意，可供他人使用。」至於文本處理能力描述則為「能對長篇且具難度的文本作摘要。」(Council of Europe 2001)。由此可知，將長篇文本濃縮為簡明扼要的摘要，為高程度教育領域與職場領域中不可或缺的寫作能力，因此如何評量高程度華語文學習者的摘要寫作能力為重要課題。

基於此，本研究將探討流利精通級華語文寫作測驗摘要題型的信度與評分規準的適切性。具體而言，乃分析所有評分者的評分一致性，以及探討分析式評分規準中的評分細項能否預測受試者寫作整體表現，以作為修訂評分規準之依據。希冀本研究結果能提供測驗實務界提升寫作測驗信度之參考。

2. 文獻探討

本研究之文獻探討分為四個部分：第一部分說明寫作能力的定義與評量重點；第二部分介紹流利精通級華語文寫作測驗；第三部分探討寫作測驗的信度；第四部分介紹寫作測驗評量要素，包括：評分規準、評分方式、評分流程。

2.1 寫作能力的定義與評量重點

CEFR 在寫作能力描述方面，將寫作活動與策略分為寫作表達、寫作互動與寫作文本形式三大類。寫作表達類包含創作性寫作(Creative Writing)與報告及論文(Reports and Essays)；寫作互動類包括通信(Correspondence)與便條、留言、表格(Notes, Messages and Forms)；寫作文本形式類包括記筆記與寫作文本處理。其中，與本研究最為相關的寫作能力是記筆記及寫作文本處理。包含 C1 與 C2 等級之精熟使用者(proficient user)記筆記能力描述如下：C1 為「在聆聽自己感興趣的演說時，能詳盡且正確地記筆記，且貼近講者的原意，可供他人使用。」；C2 為「能意識到演說內容的言外之意，並記下原文及引申義。」；精熟使用者之寫作文本處理能力描述如下：C1 為「能對長篇且具難度的文本作摘要。」；C2 為「能對不同來源的資訊作摘要，重新組織論點和敘述方式，以寫出語意連貫的文本。」(Council of Europe 2001)。這些能力皆是高程度教育領域與職場領域中非常重要的寫作技能。

關於與寫作相關的語言知識，Grabe 和 Kaplan（1996）認為可分為語言的知識(Linguistic knowledge)、篇章的知識(Discourse knowledge)與社交語言的知識(Sociolinguistic knowledge)。語言的知識包括寫作編碼的知識、音韻學與形態學的知識、字彙、句法結構的語言、跨語言間差異的覺察、不同語言精熟的覺察等；篇章的知識包括句中與句間的策略、訊息架構知識、原因間語意關聯的知識、組織基模知識、跨語言間，不同層次的篇章技能精熟之覺察等；社交語言的知識則包括寫作語言的實用性使用、寫作者與情境參數（如：語言使用的精熟度、涉入的程度，即保持中立抑或介入、題目的交互作用等）、對寫作者與情境參數角色的自我覺察等。以上語言知識成分，皆可作為制訂寫作評分項目的參考。另一方面，關於文句的連結與組織，Sullivan（1980）指出，段落寫作(paragraph writing)的連結必須有邏輯性、一貫性(unity)、連貫性(coherence)與持續性(continuity)，讓整個段落具流暢性(smoothness)與適切性。Brookes 和 Grundy（1998）認為，在寫作中，能將雜亂無章的材料組織成有意義且連貫的文章，是很重要的能力。張玉茹（2004）認為撰寫完整的段落是創作完整作文的開始。由此可知，形式銜接度、語意連貫性及結構完整性，是篇章表現好壞的重要因素。

2.2 流利精通級華語文寫作測驗

華語文寫作測驗(TOCFL Writing Test)為國家華語測驗推動工作委員會（簡稱華測會）專為母語非華語者所研發的標準化語言能力測驗，參照 CEFR 將測驗架構分為三等六級，三等分別為入門基礎級、進階高階級與流利精通級，而每一等級經過

標準設定(standard setting)程序，可依據測驗成績再細分成兩級，依照通過等級由低至高依序為入門級、基礎級、進階級、高階級、流利級、精通級共六級，可完整對應到 CEFR 的 A1 至 C2 等級（國家華語測驗推動工作委員會 2015, 2016, 2018）。施測方式採電腦化測驗，試題透過電腦螢幕呈現，應試者利用鍵盤輸入文字撰寫文章。

流利精通級寫作測驗一共有兩個題型，分別是摘要寫作與觀點論述。由於本研究主要探討高程度華語學習者之摘要寫作能力評量相關內容，故僅針對摘要寫作題型之信度進行探討。此題型要求應試者在 50 分鐘的測驗時間內，閱讀一篇 1000 字左右以多輪對話形式呈現的訪談資料[1]，然後再以自己的話語重新組織內容，完成 200 至 300 字的摘要。

2.3 寫作測驗信度

信度在測量領域的意義等同一致性，為測量結果不受測量誤差影響的程度（王文中等 2008）。影響測驗信度的因素包括：測驗本身、受試者與評分者、測驗情境、信度評估方式等。一份測驗，倘若無論何時、何地，由任何人進行施測、計分，都能得到一致性很高的測驗結果，就表示此一測驗具有較高的信度（陳柏熹 2011）。由於寫作測驗的成績需仰賴評分者的判斷，能力評估性質屬主觀式評分，而評分不一致是造成評分錯誤的來源之一(Bachman 1990)。

常見的測驗信度評估指標包括：再測信度、複本信度、內部一致性信度、評分者信度。一般而言，寫作測驗較為著重後兩者（王德蕙等 2013）。其中，評分者信度主要是了解由不同評分者進行測驗評分時，其結果是否一致。依其性質，又分為評分者內一致性(intra-rater consistency)與評分者間一致性(inter-rater consistency)，前者是指同一評分者在給分上的一致性或穩定性；後者則是指不同評分者在評量相同受試者時，其評量分數或分數等級的一致性（陳柏熹 2011）。針對單一評分者，需檢視評分者內信度(Bachman 1990)；若由多位評分者進行評分，則可運用類推性理論（generalizability theory，簡稱 G 理論）進行評分者間信度的估計（張郁雯 2009）。

類推性理論是估計透過測驗、評定量表、問卷或觀察表等程序所獲得測量結果的可靠性(dependability)。此理論延伸並擴展古典測驗理論對於信度及測量誤差來源的概念，並應用變異數分析的方法進行分析。類推性理論能確認並估計多種不同誤差來源的變異數，以及估計不同來源之間交互作用的變異數。此外，也能預測當評量人數、試題數量或情境改變等等時，測量信度改善的情形(Shavelson, Webb and Rowley 1989; Mushquash and O'Connor 2006; Alkharusi 2012)。相較於古典測驗理論的

[1] 摘要寫作例題可參見 https://www.sc-top.org.tw/chinese/WT/bandC-3.php

信度測量方式，一次只能考量一個誤差來源，單一種信度估計結果無法類推到不同的信度估計方式，即便獲得不同信度估計方式的結果，也難以統整訊息以進行決策(Webb, Rowley and Shavelson 1988)，類推性理論則提供了更有彈性和實用性的架構，能同時估計多種誤差來源的影響。

類推性理論分為類推性研究(generalizability study, G study)和決策性研究(decision study, D study)，類推性研究估計進行測量時潛在測量誤差來源的變異量，決策性研究則為了特定目的，應用類推性研究的分析結果來設計測量方式以讓誤差最小化(Shavelson et al. 1989)。在類推性理論中，測量程序的可靠性以 G 係數(G coefficient)表示。G 係數可再分為相對 G 係數(relative G coefficient)和絕對 G 係數(absolute G coefficient)，前者類似於古典測驗理論的信度係數，也被稱為類推性係數，後者有時又被稱為 phi 係數或可靠性係數(Brennan 2000; Mushquash and O'Connor 2006)。當結果用途為常模參照時，較適用類推性係數；如用於標準參照解釋時，可靠性係數較適用(Webb et al. 1988)。G 係數介於 0 至 1 之間，當係數越高時，表示研究者越能將獲得的分數類推到研究其他面向。Shavelson 和 Webb（1991）認為，係數的數值應高於.80，測驗結果才可靠。

以計算評分者間相關的方法來說，當評分者間的評分變異小，所得的相關係數可能反而變小，而導致錯誤的研究結果（王德蕙等 2013）。許多研究結果顯示，實作評量特別容易受到評分者偏誤、計分主觀性及受試者因素的影響，類推性理論最適合用來估計此類評量的信度（張郁雯 2009），比傳統的信度驗證有更多的優點(Kretchmar 2006)，若能以類推性理論為主，輔以評分一致率與評分者間相關等數據，將使測驗分數的信度證據之驗證更加完整（王德蕙等 2013）。然而以類推性研究評估寫作能力的研究相當罕見(Chen, Niemi, Wang, Wang and Mirocha 2007)，實屬可惜。

2.4 寫作測驗評量要素

Alderson、Clapham 和 Wall（1995）認為在主觀性評分的流程中，評分領導者與評分者需注意以下事項，方能確保評分信度：在評分領導者方面，應在評分前確定評分規準；在評分者方面，必須預先接受評分訓練，除了要熟悉評分規準以外，還需要了解評分方式與評分流程等相關內容，並經過試評，確定其評分標準與評分規準的描述一致之後，再進行正式評分（Weigle 2002；陳柏熹 2011；熊玉雯、李慧萱、宋曜廷 2014）。換言之，評分規準、評分方式、評分流程三者，與評分信度息息相關，茲分述如下。

2.4.1 評分規準

評分規準是評分者評定寫作測驗受試者作答反應分數等級之依據（陳柏熹 2011）。Bachman（1990）認為，評分者在評閱不同受試者文本時，若能持續依循同一套評分標準，則其評分結果是可信的。Mathews（1985）亦指出，若不同評分者對於評分規準的解讀和應用有所差異，可能影響測驗的公正性。至於評分規準的制定方法，則必須考量若干要素，如：評分項目與評分量尺(rating scale)等等（引自 Alderson et al. 1995）。過去也有學者提出評分者能否有效區分各個分項的疑慮，對此，Luoma（2004）由認知心理學的角度指出，4 至 5 個評分項目是認知能力所能負荷之範圍。至於評分量尺的制定，則應避免使用過於簡略的敘述，以免不同評分者的解讀有所差異，也不宜超過 7 個等級，以免各等級之間的差異過於細微(Alderson et al. 1995)。

國際間各大型測驗單位也分別依據其測驗特性制定不同的評分規準，以下分別介紹若干英語檢定與中文檢定的寫作測驗評分規準。在英語檢定部分，如：（1）劍橋國際英語認證(Cambridge English Language Assessment[2])之 Proficiency(CPE)等級寫作測驗的評分級距為 0-5 級分，評分項目包括：內容(Content)、溝通完成度(Communicative Achievement)、組織(Organisation)、及語言(Language)。（2）多益英語測驗(TOEIC[3])之寫作測驗由描述圖片、回覆書信要求、陳述意見三種題型組成。其中，難度最高的題型是陳述意見，其評分級距為 0-5 級分，評分項目包括對主題與任務的闡述、語法、詞彙與組織。（3）全民英語檢定(GEPT[4])分為 5 個等級，其優級寫作測驗採用分項式評分量表，評分項目包括內容、組織、用字遣詞、語法結構、作者必須傳達給讀者的訊息及作者角色等，量表中對各評分項目的通過標準(passing criteria)皆予以說明。

在中文檢定部分，如：美國外語教學委員會(ACTFL)寫作測驗的評分規準架構包含 4 大面向，分別是目的內容、組織結構、語法句法與詞彙標點（熊玉雯等 2014）。北京語言大學於 2006 年所推出漢語水平考試(HSK)改進版，在初、中、高三個級別的考試中都設計了寫作測驗，而高級寫作主要考查撰寫結構完整的議論性短文的能力，要求語句準確、表達連貫與得體、結構完整、條理清晰、漢字書寫清楚工整、標點符號準確，篇幅 400-600 字。評分方法以綜合評分法為主，輔以分項評分標準進行等級描述，評分標準可概括出「內容」、「語言形式」、「語篇表達」三大要素。

[2] 取自 https://www.cambridgeenglish.org/exams-and-tests/proficiency/
[3] 取自 http://www.toeic.com.tw/sw/file/TOEIC_Speaking_and_Writing_Examinee_Handbook.pdf
[4] 取自 https://www.gept.org.tw/Exam_Intro/t05_introduction.asp

內容項目強調議論性，主要看論點是否切題，論據與論述是否充分；語言形式項目包括詞句等的規範性與多樣性；語篇表達項目強調連貫性，包含語言形式和內在邏輯的連貫性（聶丹 2009）。

由上可知，不同寫作測驗單位或因語種特性或是題型因素，對於評量項目的側重點與詮釋有所差異，然而評量要項不外乎內容、結構組織與語言能力，其他尚有錯別字、規範性、標點符號及篇幅等。以對內容的詮釋為例，有的是指對主題與任務的闡述，有的則是指立意取材，儘管詮釋不同，其實內涵大抵相同。至於華測會研發之華語文寫作測驗摘要寫作評分規準乃參考學者專家對於評分量尺制定上的建議、各大寫作測驗單位的評量要項，及摘要寫作之特性研製而成，評分向度分為「任務完成度」與「語言表現」，後者又分為「句型詞藻表現力」與「詞彙語法正確度」兩個評分子項。任務完成度主要檢視原文之重點擷取及組織架構表現；句型詞藻表現力檢視能否展現高程度句型結構與詞語的運用能力；詞彙語法正確度則是檢視詞彙語法掌握度。

關於評分規準合宜性的驗證方法，有些學者提出可透過多元迴歸分析方法進行檢視，因多元迴歸分析方法是以多個預測變項預測一個效標變項，分析各個預測變項對於效標變項影響的程度；同時具備篩選預測變項的能力，從而發展、檢定多個包含不同預測變項的模式。其中的預測變項對於效標變項的預測力，意即解釋變異量，可作為檢測寫作測驗評分規準合宜性的指標（王德蕙等 2013）。

2.4.2 評分方式

寫作測驗評分方式，一般分為整體式評分(holistic scoring)與分析式評分(analytic scoring)。前者是對受試者作答內容的各方面表現做整體性考量之後，給予單一分數。後者則是根據多個分項的描述，各給予一個分數，再將各分項分數依計分規定加總或平均，得到一個最後的分數（陳柏熹 2011）。雖然分析式計分較耗時費事，但在建構效度(construct validity)上能夠反映出學習者分項的能力發展，且在後效作用(impact)上，不僅能提供學習者診斷性的回饋，並能提供評分者更有效的訓練機制（熊玉雯等 2014）。由於限制反應型申論題有比較明確的答題方向與內容，因此可以根據這些方向與內容，採用分析式評分來進行評分（陳柏熹 2011）。對於此兩種評分方式的差異，Weigle（2002）認為分析式評分的信度比整體式高，也因為不同面向寫作能力的發展程度可能有所不同，因此更適合二語學習者，亦能提供更多有用的診斷資訊，且更適合用以訓練評分者，然而就效率而言，整體式評分方式相對快速且簡單。

2.4.3 評分流程

Bachman（1990）指出，除了是否持續依循同一套評分規準之外，評分流程也可能影響評分一致性。對於如何制定標準化評分流程，Alderson 等人（1995）認為在正式評分之前，評分領導者應快速並大量地閱讀文本，以掌握受試者的寫作模式，以及在寫作過程中遭遇的問題。同時，也必須挑選出各個級分的範文，以便進行討論，並藉此修訂評分規準，最後提供評分者作為評分依據。且所有評分者必須定期接受訓練，以免產生個人的評分模式，同時建議在評閱的過程中，帶領者需不時地檢查評分者的評閱狀況，以確保他們維持標準。此外，帶領者必須給予評分者評價並記錄其表現，對於不適任者，依狀況再次給予訓練或終止其評分資格。

綜合學者專家的研究成果可知，測驗本身、受試者與評分者、測驗評估方式等，皆為影響測驗信度的重要因素。對寫作測驗而言，評分不一致是造成評分錯誤的來源之一，當由多位評分者進行評分時，若採用類推性理論進行評分者間信度的估計，可推估不同來源誤差的變異量，亦有助於預測評分者人數對信度的影響。在制定評分規準方面，有學者提出 4 至 5 個評分項目是認知能力能負荷的範圍，評分量尺則不宜超過 7 個評分級距，在評量內容部分，顯示不同測驗單位對此看法大同小異。除此之外，可透過迴歸分析方法檢驗評分規準的合宜性。在評分方式上，不少學者認為分析式評分方式雖較耗時費事，但有助於測驗單位提供評分者更有效的評分培訓機制。

從過去的研究成果來看，學者專家對於華語文寫作測驗分析式評分規準的研究近乎闕如。緣此，本研究除了對流利精通級華語文寫作測驗摘要寫作題型進行信度分析之外，亦探討評分規準的適切性。

3. 研究方法

3.1 研究樣本與評分者

華測會於 2016 年 11 月 3 日舉辦第一次流利精通級華語文寫作測驗，研究樣本為參加該次測驗的 57 名受試者的摘要寫作文本。評分者一共有 6 位，其中 1 位是華測會寫作測驗研發人員，另外 5 位評分者的華語教學年資皆介於 10 至 15 年，且皆接受過華測會 4 場以上分析式評分培訓會議。

3.2 評分規準與評分模式

為了探討評分規準的適切性，本研究分別於 2016 年 11 月與 2018 年 1 月進行評分作業，第一次採取分析式評分法，第二次採取整體式評分法，兩次的參與者皆相

同。採取分析式評分法時，6 位評分者均評閱 57 篇文本；採整體式評分法時，因挑選出 6 篇作為說明級分樣卷，故 6 位評分者評閱篇數皆為 51 篇。分析式評分作業結束後，華測會研發人員彙整所有評分資料，如遇評分者間給分差距超過 1 級分、無法計算出眾數或眾數不只一個時，則進行內部討論取得共識，以確定最後成績。以下說明這兩次評分作業所使用的評分規準與評分模式。

3.2.1 分析式評分規準與評分模式

此部分先介紹流利精通級華語文寫作測驗摘要題型分析式評分規準的架構，包含評分級距與評分項目，再介紹分析式評分法之評分模式與分數轉換。

3.2.1.1 分析式評分規準的評分級距與評分項目

分析式評分規準的評分級距為 0-5 級分，倘若文本出現如：未達 140 個字、完全抄錄原文、文不對題、不知所云或未以中文寫作等情況中的任何一項，將被評為 0 級分，1 級分表示該文本的寫作表現未達流利級，2 至 3 級分屬流利級，4 至 5 級分屬精通級。評分向度分為「任務完成度」與「語言表現」，向度的評分級距為 0-5 級分。任務完成度向度包含「訊息」、「組織」與「規範性」3 個評分細項，分數範圍皆為 0-3 分。語言表現向度又分為「句型詞藻表現力」與「詞彙語法正確度」兩個評分子項，評分級距皆為 0-5 級分。句型詞藻表現力包含「複雜句型」、「多樣句型」、「詞語運用」與「簡潔性」4 個評分細項，分數範圍皆為 0-3 分；詞彙語法正確度包含「詞彙語法」、「漢字運用」、「標點符號」與「分行分段」4 個評分細項，前兩項的分數範圍為 0-3 分，後兩項則是 0-2 分。亦即摘要寫作分析式評分規準一共有 11 個評分細項，所有的評分細項與分數範圍，如表 1 所示。

表 1：分析式評分規準的評分細項與分數範圍

評分向度	評分子項	評分細項	分數範圍
任務完成度	--	訊息	0-3
		組織	0-3
		規範性	0-3
語言表現	句型詞藻表現力	複雜句型	0-3
		多樣句型	0-3
		詞語運用	0-3
		簡潔性	0-3

表 1：分析式評分規準的評分細項與分數範圍（續）

評分向度	評分子項	評分細項	分數範圍
語言表現	詞彙語法正確度	詞彙語法	0-3
		漢字運用	0-3
		標點符號	0-2
		分行分段	0-2

3.2.1.2 分析式評分法的評分模式與分數轉換

分析式評分法之評分模式，是由評分者依據受試者在 11 個評分細項上的表現，在華測會寫作測驗線上評分系統上，分別勾選相應分數。各細項分數傳輸至評分系統後，便自動按照華測會既定的成績轉換方式產出子項級分、向度級分與整體級分。以下分別說明這三類級分的產出方式。

首先說明子項級分的產出方式：評分子項的評分級距是 0-5 級分，成績是由該子項的所有評分細項的分數轉換而來。以句型詞藻表現力為例，若在 4 個評分細項（複雜句型、多樣句型、詞語運用、簡潔性）中，有 3 個細項得 3 分，1 個細項得 2 分，則該子項的成績為 4 級分。句型詞藻表現力之子項級分與細項分數之間的轉換方式，如表 2 所示。

表 2：句型詞藻表現力之子項級分與細項分數之間的轉換方式

子項級分	複雜句型、多樣句型、詞語運用、簡潔性之分數的分布狀況
5 級分	4 個皆 3 分
4 級分	3 個 3 分，1 個 2 分
3 級分	2 個 3 分，2 個 2 分
2 級分	1 個 3 分，3 個 2 分；4 個皆 2 分
1 級分	其中 1 個 1 分
0 級分	符合 0 級分特徵中任何 1 項

詞彙語法正確度之子項級分與細項分數之間的轉換方式，如表 3 所示。以 4 級分為例，若詞彙語法得 3 分、漢字運用得 2 分、標點符號與分行分段都得 2 分（滿分），則此一子項得 4 級分。

表 3：詞彙語法正確度之子項級分與細項分數之間的轉換方式

子項級分	詞彙語法、漢字運用、標點符號、分行分段之分數的分布狀況
5 級分	詞彙語法 3 分；漢字運用 3 分；標點符號 2 分；分行分段 2 分
4 級分	詞彙語法 3 分；漢字運用 2 分；標點符號 2 分；分行分段 2 分
3 級分	詞彙語法 2 分；漢字運用 3 分；標點符號 2 分；分行分段 2 分
2 級分	4 個皆 2 分
1 級分	其中 1 個 1 分
0 級分	符合 0 級分特徵中任何 1 項

其次說明向度級分的產出方式：任務完成度向度級分是從它的 3 個細項分數轉換而來；語言表現向度級分則是從它的 2 個子項級分轉換而來。任務完成度向度級分與細項分數之間的轉換方式如表 4 所示。以 2 級分為例，若訊息或組織之中有一項得 3 分，其他皆 2 分，或是 3 個細項皆得 2 分，其向度級分皆為 2 級分。

表 4：任務完成度向度級分與細項分數之間的轉換方式

向度級分	訊息、組織、規範性之分數的分布狀況
5 級分	3 個皆 3 分
4 級分	訊息或組織其中一項 2 分，其他皆 3 分
3 級分	規範性 2 分，其他皆 3 分；規範性 3 分，其他皆 2 分
2 級分	訊息或組織其中一項 3 分，其他皆 2 分；3 個皆 2 分
1 級分	其中 1 個 1 分
0 級分	符合 0 級分特徵中任何 1 項

語言表現向度級分與子項級分的轉換方式是依以下 2 個規定計分：（1）若兩個子項的級分相差 1 個級分，以低分者為向度級分。（2）若兩個子項的級分相差 2 級分（含以上），向度級分由低者往上加一個級分。語言向度級分與子項級分之間的轉換方式，如表 5 所示。

表 5：語言向度級分與子項級分之間的轉換方式

向度級分	句型詞藻表現力、詞彙語法正確度之級分的分布狀況
5 級分	2 個皆 5 級分
4 級分	2 個皆 4 級分；1 個 5 級分、1 個 4 級分；1 個 3 級分、1 個 5 級分
3 級分	2 個皆 3 級分；1 個 3 級分、1 個 4 級分；1 個 2 級分、1 個 4 級分；1 個 2 級分、1 個 5 級分
2 級分	2 個皆 2 級分；1 個 2 級分、1 個 3 級分；1 個 1 級分、1 個 3 至 5 級分
1 級分	2 個皆 1 級分；1 個 1 級分、1 個 2 級分
0 級分	符合 0 級分特徵中任何 1 項

最後說明整體級分的產出方式：整體級分的評分級距為 0-5 級分，其 2 級分為流利級通過門檻。整體級分是由任務完成度與語言表現兩大向度的級分轉換而來，轉換方式是依以下 2 個規定計分：（1）在兩個向度的級分皆高於 1 級分的情況下，若兩者相差 1 級分，則取低者；若相差 2 級分（含以上），則是由低者往上加一個級分。（2）只要其中一個向度的級分為 1 級分，整體級分為 1 級分。之所以採取較為嚴格的評定標準，乃基於流利精通級為寫作測驗最高等級，惟有各個面向皆達到一定水準，才能展現此一等級該具備的寫作表現。向度級分與整體級分之間的轉換方式，如表 6 所示。

表 6：向度級分與整體級分之間的轉換方式

整體級分	任務完成度、語言表現之級分的分布狀況
5 級分	2 個皆 5 級分
4 級分	2 個皆 4 級分；1 個 5 級分、1 個 3 或 4 級分
3 級分	2 個皆 3 級分；1 個 5 級分、1 個 2 級分；1 個 4 級分、1 個 3 或 2 級分
2 級分	2 個皆 2 級分；1 個 3 級分、1 個 2 級分
1 級分	其中 1 個 1 級分
0 級分	符合 0 級分特徵中任何 1 項

3.2.2 整體式評分的評分規準與評分模式

整體式評分所使用的評分規準，在評分級距與評量重點上，跟分析式評分的評

分規準並無二致，惟呈現的形式不同，整體式評分原則詳見表 7 所示。在整體式評分過程中，評分者不需對評量要項逐一給分，而是綜合考量受試者在任務完成度、句型詞藻表現力與詞彙語法正確度各方面的表現，給予一個 0-5 級分的整體評分。

表 7：摘要寫作題型整體式評分原則

級分	評分規準	
5	任務完成度	訊息近乎完整，且組織良好。
	句型詞藻表現力	能運用複雜句型，句型靈活多變；能將淺白用語轉換為高程度用語，容許極少數未轉換，極少數冗贅重複。
	詞彙語法正確度	容許極少數詞彙語法錯誤，容許極少數增字／漏字／錯別字。
4	任務完成度	訊息近乎完整，組織大致良好，或是訊息大致完整，組織良好。
	句型詞藻表現力	能運用複雜句型，句型多樣；能將淺白用語轉換為高程度用語，容許極少數未轉換，極少數冗贅重複。
	詞彙語法正確度	容許極少數詞彙語法錯誤，容許少數增字／漏字／錯別字。
3	任務完成度	訊息近乎完整，組織良好，但連續數句近乎抄錄原文，或加上少部分作者的觀點。 訊息大致完整，組織大致良好。
	句型詞藻表現力	句型多樣；能將淺白用語轉換為一般書面用語，容許極少數未轉換，極少數冗贅重複。
	詞彙語法正確度	容許少數詞彙語法錯誤，容許極少數增字／漏字／錯別字。
2	任務完成度	訊息近乎完整，組織大致良好，但連續數句近乎抄錄原文或是加上少部分作者的觀點。 訊息大致完整，組織良好，但連續數句近乎抄錄原文或是加上少部分作者的觀點。 訊息大致完整，組織大致良好，但連續數句近乎抄錄原文或是加上少部分作者的觀點。

表 7：摘要寫作題型整體式評分原則（續）

級分	評分規準	
2	句型詞藻表現力	句型多樣；能將淺白用語轉換為一般書面用語、容許少數未轉換、少數冗贅重複。
	詞彙語法正確度	容許少數詞彙語法錯誤、容許少數增字／漏字／錯別字。
1	任務完成度	訊息不足、組織不佳、大量抄錄原文、加上過多作者的觀點、以第一人稱撰寫、超過 400 字。
	句型詞藻表現力	句型結構過於簡單或變化性不足或多處不適切；詞語過於淺白／口語或冗贅重複多。
	詞彙語法正確度	詞彙語法錯誤多；增字／漏字／錯別字多；過度任意分行分段；標點符號錯誤多。
0	未達 140 字、完全抄錄原文、文不對題、不知所云、未以中文寫作。	

3.3 資料分析

本研究透過斯皮爾曼等級相關、類推性理論、多元迴歸分析三種分析模式探討評分信度，使用分析軟體為 SPSS 23.0 版。以下介紹此三種分析模式在本研究中的運用方式。

3.3.1 斯皮爾曼等級相關

此一方法的分析內容分為兩類：（1）將所有評分者採取分析式評分方式所評定出的向度級分跟華測會確定的向度級分（簡稱為「最後向度成績」）進行分析。（2）將所有評分者分別以分析式評分方式、整體式評分方式所評定出的整體級分跟華測會確定的整體級分（簡稱為「最後成績」）進行分析。一般來說，相關係數達.4 以上即具有中度相關，達.7 以上表示兩者具有高度相關。

3.3.2 類推性理論

本研究針對摘要寫作題型的分析式評分結果進行探討，誤差來源為評分者，以及受試者與評分者之間的交互作用。由於 6 位評分者皆評閱所有受試者的寫作文本，故分析模式為 p×r 完全交叉設計，p 表示受試者，r 表示評分者，「×」表示交叉。使用 Mushquash 和 O’Connor（2006）撰寫的 SPSS 語法進行分析。

3.3.3 多元迴歸分析

採取多元迴歸分析方式中的逐步迴歸法(stepwise)，首先，本模式從所有預測變項中選出對整體式評分之整體級分具有預測力的變項，並針對挑選出之變項進行分析。其次，以強迫輸入法(enter)，強迫模式將所有的預測變項皆納入分析。在此，所有預測變項是指分析式評分的評分細項，依序為訊息、組織、規範性、複雜句型、多樣句型、詞語運用、簡潔性、詞彙語法、漢字運用、標點符號、分行分段，總計有 11 個評分細項，前 9 項的評分級距是 0-3 級分，後 2 項則是 0-2 級分；整體級分則是採整體式評分所得的資料。

4. 研究結果與討論

本部分說明信度分析與多元迴歸分析結果。信度分析方式採取斯皮爾曼等級相關跟類推性研究分析；多元迴歸分析是以逐步迴歸法(stepwise)與輸入法(enter)檢視變項對整體式評分之整體級分的預測力。

4.1 信度分析

4.1.1 斯皮爾曼等級相關分析

此一部分將探討所有評分者採取分析式評分方式時，所評定出的向度級分跟最後向度成績的一致性，以及所有評分者分別以分析式評分方式、整體式評分方式所評定出的整體級分跟最後成績的一致性。

4.1.1.1 評分者的向度級分與最後向度成績的一致性分析

整體級分係由任務完成度與語言表現兩大向度的級分轉換而來，故先分析評分者所評定之向度級分與最後向度成績的等級相關，其結果如表 8 所示。在任務完成度向度部分，除了 C01、C02 的相關係數未達.70，表示評分品質需加強之外，其餘 4 位的相關係數介於.702 至.939 之間，顯示評分結果具可信度；在語言表現向度部分，仍是 C06 的相關係數最高，相關數值為.880，顯示具良好的信度。反之，相關數值最低的是 C02，僅達.44，檢視其語言表現向度給分的標準差，相較於其他評分者及最後向度成績，標準差數值較小，給分較為集中，表示評分品質有待加強。未來可透過更多的討論以釐清其評分盲點，並給予具體的評分建議，亦需持續性地觀察評分狀況有無改善。

表 8：評分者的向度級分與最後向度成績之斯皮爾曼等級相關（N = 57）

評分者 / 評分向度	C01	C02	C03	C04	C05	C06
任務完成度	.671**	.606**	.836**	.702**	.863**	.939**
語言表現	.755**	.440**	.715**	.792**	.818**	.880**

註：** $p < .01$

4.1.1.2 評分者分析式和整體式評分之整體級分與最後成績的一致性分析

以分析式評分方式評閱的文本篇數為 57 篇，以整體式評分方式評閱時，因需從中選取 6 篇作為說明樣卷，因此評閱篇數為 51 篇。所有評分者分別以分析式評分方式與整體式評分方式所評定出之整體級分，跟最後成績的斯皮爾曼等級相關分析結果，如表 9 所示。在採取分析式評分方式的情況下，相關係數最高的是 C04 與 C06，最低是 C01。所有評分者的相關係數介於.718 至.913 之間，顯示評分結果具可信度。換言之，所有評分者的整體級分與最後成績的斯皮爾曼等級相關呈中度至高度正相關，達.01 之顯著水準，整體而言，評分品質良好。而在採用整體式評分方式的情況下，由於 C06 是華測會研發人員因此相關係數較高為.962 之外，其餘 5 位的相關係數介於.567 至.691 之間（$p < .01$），與最後成績的相關係數均低於.70。推測其原因是因首次採取整體式評分，評分者尚未習慣同時考量所有評量要點之評分方式，僅憑各自著重的面向給分；抑或是需要同時全面考量所有評分項目並整合成單一分數，對評分者的認知負荷量較重，以致與最後成績的相關係數偏低。

表 9：評分者分析式和整體式評分之整體級分跟最後成績的斯皮爾曼等級相關

評分者 / 評分方式	C01	C02	C03	C04	C05	C06
分析式評分	.718**	.765**	.852**	.913**	.851**	.913**
整體式評分	.690**	.624**	.691**	.624**	.567**	.962**

註：** $p < .01$

4.1.2 類推性研究分析

由於評分者間相關的信度估計方式在評分者間的給分一致性高時，分析所得的相關係數反而變小，可能導致錯誤的研究結果，且無法進一步得知評分人數的多寡對測驗分數信度的影響（王德蕙等 2013），若同時進行斯皮爾曼等級相關分析與類推性研究分析，則能得到更多資訊。本研究透過類推性研究得出評分信度，如表 10

所示。

以下將依序說明整體級分、任務完成度向度、語言表現向度在受試者與評分者的變異成分差別。整體級分的受試者變異量為 1.445，變異成分達 76.5%，而評分者的變異量僅為 0.012，變異成分僅占 0.6%；任務完成度向度的受試者變異量為 1.869，變異成分達 68.1%，而評分者的變異量僅為 0.038，變異成分僅占 1.4%；語言表現向度的受試者變異量為 1.126，變異成分達 62.7%，而評分者的變異量僅為 0.013，變異成分僅占 0.7%。由此可見，不論從整體級分或是從兩大向度級分來看，受試者的變異成分皆最高，交互作用及殘差的變異成分居次，變異成分依序為 22.8%、30.5%、36.6%，而評分者的變異成分最低，其結果顯示受試者的得分能反映其真實寫作能力。同時顯示評分結果可以涵括受試者在任務完成度與語言表現這兩個評分向度上的寫作能力，而非僅著重某一向度。

表 10：類推性研究分析結果

	整體級分		任務完成度向度		語言表現向度	
變異成分	變異量	百分比	變異量	百分比	變異量	百分比
受試者	1.445	76.5%	1.869	68.1%	1.126	62.7%
評分者	0.012	0.6%	0.038	1.4%	0.013	0.7%
交互作用及殘差	0.430	22.8%	0.838	30.5%	0.656	36.6%
總變異	1.887	100.0%	2.745	100.0%	1.795	100.0%

評分者人數跟可靠性數值的關係如圖 1 所示。當評分人數由 1 人增加到 2 人時，整體級分、任務完成度向度、語言表現向度的可靠性係數增加的幅度最大，其中，整體級分增加了.10；任務完成度向度增加了.13；語言表現向度增加了.14。但隨著人數增加，可靠性係數值增加的幅度愈不明顯。且若是 1 人評閱，在整體級分、任務完成度向度、語言表現向度三方面的可靠性係數，皆未達 Shavelson 和 Webb（1991）所提出的.80，顯示評分結果的可信度皆不足；若是 2 人評閱，前兩者雖超過.80，但語言表現向度為.77，略低於標準值；若是 3 人評閱，則是兩大向度皆超過.80，整體級分更高達.91。考量華語文能力測驗為一高風險測驗，對於信度應有較高標準之要求，因此一篇文本由 3 位評分者評閱，在整體級分與兩個向度級分均達.80 以上，相對經濟且具有良好的信度。

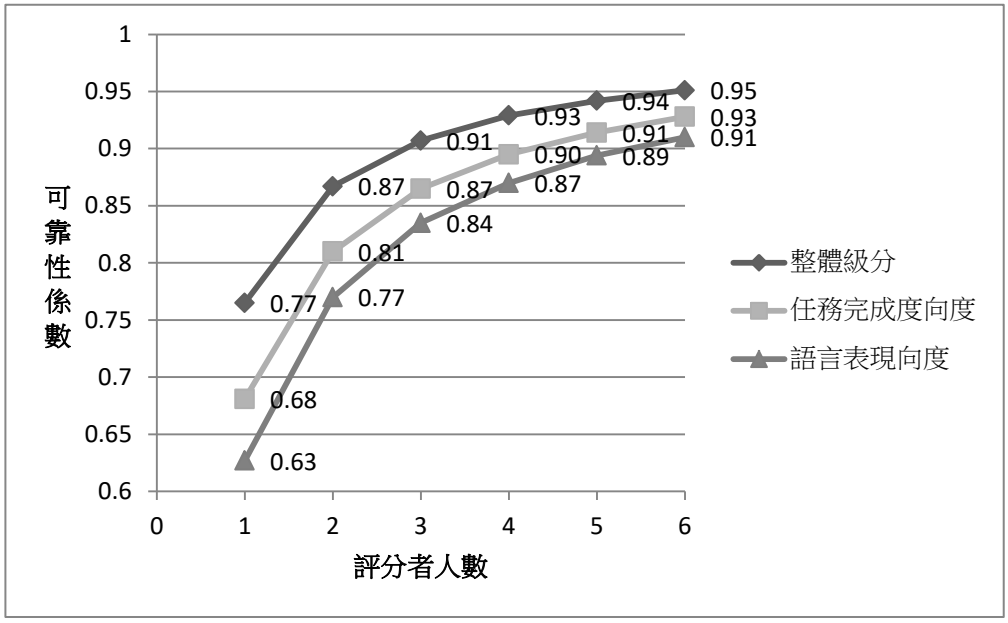

圖 1：評分者人數與可靠性係數值的關係

4.2 多元迴歸分析結果

分析式評分所使用的評分規準，包含任務完成度與語言表現兩大評分向度，前者有 3 個評分細項，後者則有 8 個，亦即評分者需要根據各個評分細項的寫作表現評定出 11 個成績，而整體式評分則是綜合考量任務完成度與語言表現兩大向度的表現，給予一個 0-5 級分的整體評分。本部分的多元迴歸分析，先採取逐步迴歸法，以分析式評分所得的 11 個評分細項的成績為預測變項，以整體式評分所得之整體級分為效標變項，其分析結果如表 11 所示。此一模式從 11 個預測變項中，選出對整體式評分之整體級分較有預測力的 5 個變項，分別是複雜句型、組織、詞彙語法、詞語運用、規範性，其聯合解釋變異量為 45.6%。由逐步多元迴歸分析得出預測整體式評分之整體級分的標準化迴歸方程式如公式 1 所示，其中，標準化迴歸係數 ß 的絕對值高低表示該變項對依變項（整體級分）的預測力高低。從表 11 的 ß 值可以看出，預測變項的預測力由高至低依序為複雜句型（ß = .374）、組織（ß = .310）、詞彙語法（ß = .186）、詞語運用（ß = -.161）、規範性（ß = .130）。

公式 1：整體級分$_{整體式}$ = .374X $_{複雜句型}$ + .310X $_{組織}$ + .186X $_{詞彙語法}$ + -.161X $_{詞語運用}$ + .130X $_{規範性}$

表 11：複雜句型、組織、詞彙語法、詞語運用、規範性對整體式評分之整體級分的逐步迴歸分析摘要表

次序	投入變項	R	R^2	調整後的R^2	淨 F 值	標準化迴歸係數 ß	t 值
1	複雜句型	.566	.320	.318	143.044**	.374	6.478**
2	組織	.639	.408	.404	104.485**	.310	5.797**
3	詞彙語法	.660	.436	.430	77.672**	.186	3.574**
4	詞語運用	.670	.449	.442	61.419**	-.161	-3.037**
5	規範性	.682	.464	.456	52.037**	.130	2.905**

註：** $p < 0.01$

反之，訊息、多樣句型、簡潔性、漢字運用、標點符號、分行分段這 6 個變項則未被模式選出。為了探究所有預測變項對整體式評分之整體級分的預測力，故另採輸入法(enter)，強迫模式將 11 個預測變項皆納入分析，得出聯合解釋變異量約為 45%。此一數值表示，儘管輸入法模式多分析了 6 個變項，其結果跟逐步迴歸法選出 5 個變項的分析結果近乎相同，研究者先檢視 6 個變項與其他變項間是否有過高的相關性，即存在共線性問題(collinearity)而影響模式估計迴歸係數的結果（吳明隆 2000）。然而從容忍度(tolerance)跟變異數膨脹因素(VIF)指標的數值中，並未觀察到這 6 個變項與其他變項有共線性。據此，研究者試圖從認知負荷量、研究樣本特性、各個評分細項的得分分布狀況這三個層面，來探討 6 個變項對整體級分預測力不顯著之成因，茲分述如下：

從認知負荷量的觀點來看，在整體式評分模式下，雖然要求評分者需要綜合考量各項寫作表現進行評分，但由於評分者的認知負荷量是有限的，僅能以若干評量重點來評定寫作表現的成績，難以顧及全面。此一推論可由評分結果驗證，在迴歸模式選出的 5 個變項中，組織與規範性隸屬於任務完成度向度，複雜句型和詞語運用隸屬於句型詞藻表現力子項，詞彙語法隸屬於詞彙語法正確度子項。此結果顯示評分者雖能關注到受試者在兩大向度與兩大子項的寫作表現，但仍較著重其中 5 個變項，無法兼顧所有變項。此一推論亦可由其中 5 位評分者在採取整體式評分模式下所評定的整體級分與最後成績的相關係數介於.567 至.691，而採分析式評分時的相關係數則高達.718 至.913 之結果得以印證。

再者，有關研究樣本特性，本研究採用之樣本乃首次流利精通級華語文寫作測驗的受試者文本，從考試結果可知此次受試者的寫作能力普遍較高，不少受試者在

某些變項均取得高分，已達天花板效應，使得部分變項對整體級分的預測力不顯著。若欲取得更確切的分析結果，宜持續蒐集不同程度之受試者文本，並進行後續研究。

最後觀察所有受試者在各個變項的成績分布狀況，有助於探究多元迴歸分析模式之所以選出與未選出某些變項的可能原因。在此先討論任務完成度向度，再討論語言表現向度。

任務完成度向度部分，模式選出的變項為組織與規範性，而未將訊息納入，亦即訊息被視為對整體式評分之整體成績預測力不足的變項。推測其原因，可能是因為華測會考量到不同的受試者對於在一千字左右的長篇多輪對話中，判定哪些內容為重要細節，可能會有不全然相同卻又似乎合理的詮釋，因此對此種本身較為主觀的變項採取較為寬鬆的評分標準。此外，對於流利精通級受試者而言，從內容淺白的多輪對話中擷取重點並非難事，受試者普遍較易獲取高分，因而降低了此變項的預測力，這個推測可由 80.4%受試者在此一變項得到滿分（3 分）得到印證。反觀組織的預測力高居第二（ß = .310），或許是能否善用銜接策略與銜接詞語，妥切安排訊息要點與相關細節的前後順序，以完成組織嚴謹且文意連貫的摘要，對程度偏低的受試者而言，是較難展現的寫作能力。從受試者的成績分布來看，得到 3 分的比例僅 17.0%，得到 2 分和 1 分的比例分別為 43.5%和 38.9%，可印證上述推論結果。至於規範性雖被模式選出，但其預測力並不高（ß = .130），主要是多達 80.7%的受試者得到 3 分，僅有零星受試者因違反考題規定而得低分。

語言表現向度部分，在句型詞藻表現力中，模式選出的變項為複雜句型與詞語運用。前者的預測力最高（ß = .374），可能是將若干獨立句子串連成結構較為複雜的句式並非易事，程度不夠高的受試者難以達成。由成績分布來看，32.4%受試者得到 3 分，36.9%得到 2 分，30.1%得到 1 分，顯示此變項的得分相當平均，亦即對寫作能力的判斷具有明顯的區辨性。詞語運用也有類似現象，可能也是因為能否將對話中的淺白用語轉換為具書面色彩的高程度用語此一能力，是具備一定程度的受試者才可達成的，由成績分布來看，33.7%的受試者得到 3 分，得到 2 分和 1 分的比例分別為 32.7%和 37.6%，亦相當平均。而未被模式選出的是多樣句型與簡潔性，前者主要檢視受試者能否運用多種句型，使其句型看起來靈活多變，有 20.3%的受試者得到 3 分，41.5%得到 2 分，37.6%得到 1 分，顯示較多受試者在此一變項掌握得不甚理想。至於簡潔性，則是有 62.1%的受試者得到 3 分，其原因可能是根據考試規定，受試者必須將一千字左右的文本濃縮為三百字以內的摘要，因此受試者寫出的文章多未出冗詞贅句之情況，符合文章簡潔之要求，故此一變項的預測力也不高。

詞彙語法正確度部分，只有詞彙語法被模式選出。在所有的受試者當中，有

31.7%得到 3 分，39.9%得到 2 分，27.8%得到 1 分，成績分布大致平均，顯示此一變項對於寫作能力的判斷也具區辨性。至於漢字運用、標點符號、分行分段均未被模式選出，而其成績分布皆為高分者占絕大多數。推測是因為流利精通級的受試者，華語的學習時間很長，因此對這些較為基本的要求大多已能掌握，使得這 3 個變項的預測力都不高。受試者的評分細項成績分布，如表 12 所示。

表 12：受試者的評分細項成績分布

評分細項	選入與否	成績	百分比
訊息	--	3 分	80.4%
		2 分	11.4%
		1 分	7.5%
		0 分	0.7%
組織	V（ß = .310）	3 分	17.0%
		2 分	43.5%
		1 分	38.9%
		0 分	0.7%
規範性	V（ß = .130）	3 分	80.7%
		2 分	7.2%
		1 分	11.8%
		0 分	0.3%
複雜句型	V（ß = .374）	3 分	32.4%
		2 分	36.9%
		1 分	30.1%
		0 分	0.7%
多樣句型	--	3 分	20.3%
		2 分	41.5%
		1 分	37.6%
		0 分	0.7%

表 12：受試者的評分細項成績分布（續）

評分細項	選入與否	成績	百分比
詞語運用	V（ß = -.161）	3 分	33.7%
		2 分	32.7%
		1 分	33.0%
		0 分	0.7%
簡潔性	--	3 分	62.1%
		2 分	24.8%
		1 分	12.4%
		0 分	0.7%
詞彙語法	V（ß = .186）	3 分	31.7%
		2 分	39.9%
		1 分	27.8%
		0 分	0.7%
漢字運用	--	3 分	65.0%
		2 分	24.5%
		1 分	9.8%
		0 分	0.7%
標點符號	--	2 分	89.9%
		1 分	9.5%
		0 分	0.7%
分行分段	--	2 分	92.2%
		1 分	7.2%
		0 分	0.7%

5. 結語與建議

關於流利精通級華語文寫作測驗摘要題型的信度，斯皮爾曼等級相關分析結果顯示，多數評分者在兩大向度的評分具可信度。對於評分品質有待加強的評分者，建議加強培訓，透過面對面討論以利釐清評分盲點，並持續追蹤評分狀況。從所有評分者以不同評分方式所評定出的整體級分與最後成績的斯皮爾曼等級相關分析來看，採用分析式評分方式的斯皮爾曼等級相關呈現中度至高度正相關，而採用整體

式評分方式的狀況下，大部分評分者的相關係數低於.70。推測其原因是整體式評分需同時全面考量評量重點，此超乎評分者的認知負荷度，也可能是評分者著重的評分面向有所不同。而從類推性理論分析結果顯示，受試者的變異成分最高，交互作用及殘差居次，評分者最低，此可證明測驗成績能反映受試者的寫作表現。在評分者人數跟信度係數的關係中可知，一篇文本由 3 人評閱，在實務上是相對經濟且具有良好的信度。

此外，多元迴歸分析結果顯示，在 11 個變項中，複雜句型、組織、詞彙語法、詞語運用、規範性這 5 個變項對整體式評分之整體級分較有預測力。對此，本研究從認知負荷量、研究樣本特性，以及受試者在各個變項的成績分布這三個層面來推測此分析模式選出與未選出某些變項的可能原因。儘管並非所有變項都對整體式評分之整體級分有顯著的預測力，但從學者專家的研究以及其他大型測驗單位所採取的評分架構中，可確知這些皆是寫作能力的評量要素。至於部分評分細項對於整體寫作表現的預測力未達顯著，可能與摘要寫作題型或是該場次受試者的特性有關，未來可再行評估是否需適度減少或合併評分細項，以提高評分效率。

引用文獻

Alderson, Charles J., Caroline Clapham, and Dianne Wall. 1995. *Language Test Construction and Evaluation*. Cambridge: Cambridge University Press.

Alkharusi, Hussain. 2012. Generalizability theory: An analysis of variance approach to measurement problems in educational assessment. *Journal of Studies in Education* 2.1: 184-196.

Bachman, Lyle F. 1990. *Fundamental Considerations in Language Testing*. Oxford: Oxford University Press.

Brennan, Robert L. 2000. Performance assessments from the perspective of generalizability theory. *Applied Psychological Measurement* 24.2: 339-353.

Brookes, Arthur, and Peter Grundy. 1998. *Beginning to Write*. Cambridge: Cambridge University Press.

Chang, Li-ping. 2017. The development of the test of Chinese as a foreign language (TOCFL). *Assessing Chinese as a Second Language*, eds. by Dongbo Zhang and Chin-his Lin, 21-41. Berlin: Springer.

Chen, Eva, David Niemi, Jia Wang, Haiwen Wang, and Jim Mirocha. 2007. *Examining the Generalizability of Direct Writing Assessment Tasks. Tech. Rep. No. 718*. Los Angeles,

CA: National Center for Research on Evaluation, Standards, and Student Testing, University of California.

Council of Europe. 2001. *Common European Framework of Reference for Languages: Learning, Teaching, Assessment.* Cambridge: Cambridge University Press.

Grabe, William, and Robert B. Kaplan. 1996. *Theory and Practice of Writing*. NY: Longman.

Kretchmar, Jennifer. 2006. Assessing the reliability of ratings used in undergraduate admission decisions. *Journal of College Admission* 192: 10-15.

Luoma, Sari. 2004. *Assessing Speaking.* Cambridge: Cambridge University Press.

Mushquash, Christopher, and Brian P. O'Connor. 2006. SPSS and SAS programs for generalizability theory analysis. *Behavior Research Method*s 38: 542-547.

Shavelson, Richard J., Noreen M. Webb, and Glenn L. Rowley. 1989. Generalizability theory. *American Psychologist* 44.6: 922-932.

Shavelson, Richard J., and Noreen M. Webb. 1991. *Generalizability Theory: A primer*. Newbury Park, AC: Sage.

Sullivan, Kathleen E. 1980 . *Paragraph Practice: Writing the Paragraph and the Short Composition* (4th edition). NY: MacMillan Publishing Co.

Webb, Noreen M., Glenn L. Rowley, and Richard J. Shavelson. 1988. Using generalizability theory in counseling and development. *Measurement and Evaluation in Counseling and Development* 21: 81-90.

Weigle, Sara C. 2002. *Assessing Writing.* Cambridge: Cambridge University Press.

王文中、呂金燮、吳毓瑩、張郁雯、張淑慧。2008。《教育測驗與評量－教室學習觀點》。臺北：五南出版社。[Wang, Wen-chung, Chin-hsieh Lu, Yuh-yin Wu, Yu-wen Chang, and Shu-hui Chang. 2008. *Educational Assessment: A Classroom Learning Perspective*. Taipei: Wu-Nan Book Inc.]

王德蕙、李奕璇、曾芬蘭、宋曜廷。2013。〈國民中學學生基本學力測驗寫作測驗－信度與效度分析研究〉,《測驗期刊》,第 6 卷第 1 期，151-184。[Wang, De-hui, Yi-xuan Li, Fen-lan Tseng, and Yao-ting Sung. 2013. The reliability and validity of writing assessment of competence test for junior high school students. *Psychological Testing* 6.1: 151-184.]

吳明隆。2000。《SPSS 統計運用實務》。臺北：松崗電腦圖書資料股份有限公司。[Wu, Ming-lung. 2000. *SPSS Statistical Application and Practice*. Taipei: Unalis

Corporation.]

張玉茹。2004。《國民中學學生英語寫作能力測驗的編製與其相關因素之研究》。臺北：國立臺灣師範大學博士論文。[Chang, Yu-ju. 2004. *The Study of Development of English as a Foreign Language (EFL) Writing Ability Test and Relationship with its Related Factors in Junior High School in Taiwan*. Taipei: National Taiwan Normal University Ph. D. dissertation.]

張郁雯。2009。《華語評量》。臺北：正中書局。[Chang, Yu-wen. 2009. *Chinese Language Assessment*. Taipei: Cheng Chung Book Co., Ltd.]

陳柏熹。2011。《心理與教育測驗：測驗編製理論與實務》。臺北：精策教育。[Chen, Po-hsi. 2011. *Psychological and Educational Testing: Theoretical and Practical of Test Development*. Taipei: Planned Education Ltd.]

國家華語測驗推動工作委員會。2015。《華語文能力測驗技術報告 2013-4 寫作測驗信效度》。新北市：國家華語測驗推動工作委員會。[Steering Committee for the Test Of Proficiency-Huayu. 2015. *Technical Report of TOCFL 2013-4: Reliability and Validity of the Writing Test*. New Taipei: SC-TOP.]

國家華語測驗推動工作委員會。2016。《華語文能力測驗技術報告 2014-4 寫作測驗信效度》。新北市：國家華語測驗推動工作委員會。[Steering Committee for the Test Of Proficiency-Huayu. 2016. *Technical Report of TOCFL 2014-4: Reliability and Validity of the Writing Test*. New Taipei: SC-TOP.]

國家華語測驗推動工作委員會。2018。《華語文能力測驗技術報告 2016-2 寫作測驗信效度》。新北市：國家華語測驗推動工作委員會。[Steering Committee for the Test Of Proficiency-Huayu. 2018. *Technical Report of TOCFL 2016-2: Reliability and Validity of the Writing Test*. New Taipei: SC-TOP.]

熊玉雯、李慧萱、宋曜廷。2014。〈基於 ACTFL 之華語文寫作評分規準〉，《華語文教學研究》，第 11 卷第 4 期，111-139。[Hsiung, Yu-wen, Hui-hsuan Lee, and Yao-ting Sung. 2014. Examining the ACTFL writing assessment rating scale for L2 Chinese learners. *Journal of Chinese Language Teaching* 11.4: 111-139.]

聶丹。2009。〈漢語水平考試（HSK）寫作評分標準發展概述〉，《雲南師範大學學報（對外漢語教學與研究版）》，第 7 卷第 6 期，15-20。[Nie, Dan. 2009. A historical account of the assessment criteria for HSK writing. *Journal of Yunnan Normal University (Teaching and Research on Chinese as a Foreign Language)* 7.6: 15-20.]

[審查：2019.2.21 修改：2019.3.13 接受：2019.3.20]

陳柏熹
Po-Hsi CHEN
10610 臺北市和平東路一段 162 號
國立臺灣師範大學教育心理與輔導學系、學習科學跨國頂尖研究中心
Department of Educational Psychology and Counseling,
Institute for Research Excellence in Learning Sciences
National Taiwan Normal University
No. 162, Sec. 1, Heping E. Rd., Taipei City 10610, Taiwan
chenph@ntnu.edu.tw

彭淑惠
Shu-Hui PENG
24449 新北市林口區仁愛路一段 2 號　國家華語測驗推動工作委員會
Steering Committee for the Test Of Proficiency-Huayu
No. 2, Sec. 1, Ren-ai Rd., Linkou Dist., New Taipei City 24449, Taiwan
speng0620@sc-top.org.tw

藍珮君
Pei-Jiun LAN
24449 新北市林口區仁愛路一段 2 號　國家華語測驗推動工作委員會
Steering Committee for the Test Of Proficiency-Huayu
No. 2, Sec. 1, Ren-ai Rd., Linkou Dist., New Taipei City 24449, Taiwan
martinalan@sc-top.org.tw

The Reliability and Rubric Relevance of the TOCFL Writing Test—Analyzing Band C Summary Writing as an Example

Po-Hsi CHEN
Department of Educational Psychology and Counseling,
Institute for Research Excellence in Learning Sciences
National Taiwan Normal University

Shu-Hui PENG
Steering Committee for the Test Of Proficiency-Huayu

Pei-Jiun LAN
Steering Committee for the Test Of Proficiency-Huayu

Abstract

A writing test is a form of language test that has fewer items in each test and relies heavily on subjective rating. These features of the writing test make the issue of reliability rather significant. This research focuses on the task type of Summary Writing used in the TOCFL Writing Test, Band C. By utilizing the analysis methods associated with Spearman's rank correlation coefficient, multiple regression analysis and generalizability theory, this study aims to discuss the reliability of, and the rubrics associated with, this specific task type. The results of this study can be summarized as follows: (1) By adopting an analytic rating method, the individual scores given by the majority of the raters have high-positive or medium-positive correlations with the final score. Inter-rater reliability is moderately high. (2) The variance components of the test-takers are the highest, which indicates that the scores for this task type can clearly reflect individual test-taker's writing ability. (3) Assigning three raters to rate each test essay can assure the quality and efficiency of the rating process. (4) Among the 11 variables which we have analyzed, 5 of them—complex sentence patterns, organization, vocabulary and syntax, vocabulary switching, and response requirements—can predict test-takers' writing ability more significantly.

Keywords: Band C Test, reliability, rubrics, Summary Writing, TOCFL Writing Test