華語文能力測驗技術報告—2016（2）

寫作測驗信效度

國家華語測驗推動工作委員會編著

序

中文近年來已經成為許多國家優先學習的第二外國語言，中文教學（又稱華語教學）也成為教育界新興的重要學科。世界各國的華語學習者越來越多，開始學習華語的年齡也持續下降，充分顯示華語的國際影響力與華語教學的未來發展潛力。

良好的華語教學除了要有優良的師資外，還需要仰賴優質的課程設計、教材及測驗工具。臺灣師範大學於 1956 年開始投入華語教學，這五十年來已經累積相當豐富的經驗，成為全球華語教學重鎮。我們也在教育部委託下，於 2005 年成立國家華語測驗推動工作委員會，並積極發展各式華語能力測驗，希望建立具有公信力的標準化華語測驗，成為全球知名的華語能力評量工具。

為了能有效評量出學習者的不同華語文能力水準，並且與國際的語言能力學習暨評量架構接軌，本會特別研發了不同等級的聽力、口語、閱讀以及寫作「華語能力測驗」，以及針對兒童所研發的「兒童華語文能力測驗」。測驗內容參考了歐洲共同語文參考架構，以及學習者的學習時數與華語能力發展，題材涵蓋食衣住行……等多元面向。為了讓學習華語者及華語相關領域工作者能更明白本會發展之各測驗的內容、測驗實施方式與成績公布之標準化流程，以及本年度正式考試之測驗效能評估，我們特別出版這份技術報告供大眾參考。

國家華語測驗推動工作委員會執行長　陳柏熹 謹識

目錄

表目錄

圖目錄

附件目錄

一、 前言

「華語文寫作測驗」(以下簡稱本測驗)由「國家華語測驗推動工作委員會」(以下簡稱本會)專責研發。本測驗專為母語非華語者所設計，參考「歐洲共同語文參考架構(Common European Framework of Reference for Languages，以下簡稱CEFR)」(Council of Europe，2001)，以溝通任務為導向。在命題方面，以真實情境中需要達成的各種溝通任務為設計重點；在評量方面，著重於考察受測者能否在特定語境下，藉由書面表達，有效地傳遞訊息；施測形式採電腦化測驗，試題透過螢幕呈現，受測者以鍵盤輸入文字進行寫作。

本會在臺灣地區，於 2011 年 10 月推出基礎級與進階級正式考試，因該測驗架構僅能區分受測者是否通過某一個等級，對於應試者及試務工作者而言經濟效益偏低，為能更進一步分辨通過測驗者能力的高低，同時提高測驗效能，自 2013 年起，本測驗的架構調整為三等六級，三等分別為入門基礎級、進階高階級與流利精通級，而每一等可依據測驗成績再細分為兩級，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。改版後的測驗方式(一等兩級)，應試者可依自己的學習背景或語言能力選擇範圍較廣的合適等級應考，考生即便因些微分數差距而未能通過較高等級之門檻，仍有機會通過較低等級之門檻，亦即一份測驗可同時判斷兩個等級程度。因應等級合併事宜，本會於 2013 年著手進行相關研究，並於同年 11 月推出進階高階級正式考試，2014 年 11 月推出入門基礎級正式考試，2016 年 11 月推出流利精通級正式考試。

本報告包含三個部分，首先簡介 2016 年流利精通級寫作測驗的能力描述、測驗題型、評分方式、評分原則與通過門檻；其次說明寫作測驗標準化製卷與評分作業流程；最後則是針對參加 2016 年流利精通級寫作測驗正式考試的考生資料進行分析，並評估此一年度的寫作測驗整體性信度與效度。

二、 測驗簡介

2016 年度華語文寫作測驗正式考試等級為入門基礎級(Band A)、進階高階級(Band B)與流利精通級(Band C)，依照測驗成績可細分為入門級(Level 1)、基礎級(Level 2)、進階級(Level 3)、高階級(Level 4)、流利級(Level 5)與精通級(Level 6)六個等級，分別對應 CEFR 的 A1(Breakthrough)、A2(Waystage)、B1(Threshold)、B2(Vantage)、C1(Effective operational proficiency)與 C2(Mastery)。由於 2013、2014 年寫作測驗的技術報告中已分別說明進階高階級與入門基礎級之寫作測驗相關內容(國家華語測驗推動工作委員會，2015、2016)，因此以下針對流利精通級寫作測驗之能力描述、測驗題型、評分方式、評分原則與通過門檻進行說明。

(一)能力描述

CEFR 針對語言學習者和使用者的寫作能力，設計了一份寫作能力描述總表。其中，C1 等級的學習者在寫作表達能力方面，能針對複雜的主題，撰寫清楚、完整且具有一定篇幅的文章，在文中能突顯主要議題，並論述理由及相關例證以支持其觀點，最後作出適當的結論；在寫作互動能力方面，能清楚且精確地表達自己的想法，並針對不同的對象，靈活且有效地調整書寫內容；在寫作文本處理方面，則能對長篇且具難度的文本作摘要。

C2 等級的學習者在寫作表達能力方面，則能以適當且令人印象深刻的寫作風格，與容易讓讀者抓住重點的邏輯結構，來撰寫清楚、通順且複雜的文章；在寫作互動方面，其能力與 C1 等級的學習者相同；在寫作文本處理方面，則能針對複雜的資訊，重新組織論點及敘述方式並作出語意連貫的摘要。

綜合比較以上兩個等級的寫作能力描述，可推知 C1、C2 兩個等級的學習者皆已具備相當高水準的寫作能力，而其差異主要顯現在表達及文本處理方面而非互動方面，C1 等級學習者能撰寫架構完整、說理清楚的長篇文章，精確地傳達想法，並能對較難的長篇文本作出摘要，而 C2 等級學習者則能在文章中更進一步展現文采，吸引讀者，在摘要中則能重新架構並敘寫所得資訊。本測驗參考 CEFR，並考量高程度華語寫作應具備的語言特色(以書面語為主，可能包含成語或諺語等等，口語、淺白的成分應適度減少)，制定出流利精通級寫作能力描述(此

描述依本會訂定之流利精通級寫作測驗題型分點列出，關於題型之說明，詳見本章第二部分「測驗題型」)，其內容如表 1 所示。

表 1　通過等級與能力描述

通過等級	能力描述
流利級	• 能大致完整地撰寫貼近原意的摘要，對各種主題的長篇文本資料，能大致適切地重新組織，並能大致使用多種句型與常用書面語。通篇脈絡大致清楚且文句通暢。 • 能撰寫闡述論點的長篇文章或報告，對各種議題能以清楚的邏輯結構予以評析、提出解決方案與作出結論，以表達個人觀點，並能運用多種不同的句型與常用書面語。
精通級	• 能撰寫表達原意且突顯重點的摘要，對各種主題的長篇文本資料，能適切地重新組織，並能自如地運用複雜的句型與多樣的高程度書面語。通篇脈絡清晰且文句簡練。 • 能撰寫闡述論點的長篇文章或報告，對各種議題能以縝密的邏輯結構予以評析與批判、提出完善的解決方案與作出適當的結論，以表達個人觀點，並能自如地運用多種不同的複雜句型與高程度書面語。

(二) 測驗題型

流利精通級寫作測驗乃依據 CEFR 的 C1 與 C2 之寫作能力指標設計測驗題型。CEFR 將寫作活動溝通類別分成三大類，分別為表達活動、互動活動及文本活動。其中表達活動包括「創作」、「報告」及「論文」；互動活動包括「書信」、「便條」及「留言」；文本活動包括「筆記」及「文本處理」。此三類寫作活動分別需要不同程度的寫作能力，因此測驗題型的難度也有所差異。

為了在有限的測驗時間內有效測出並區別 C1、C2 兩個等級受測者之寫作能力，本測驗研發人員(以下簡稱研發人員)根據 CEFR 對 C1 及 C2 的寫作能力描述，並評估 CEFR 各類寫作溝通活動在即席寫作測驗條件下的可行性、適切性與難度，制定出兩種題型。在上述三類寫作活動中，「互動活動」的內容通常較為簡短且生活化，而 CEFR 的 C1、C2 等級，則偏重專業職場與高等教育方面的應用能力。考量到在職場或在高等教育的課堂上，通常有撰寫摘要(作會議紀錄或課堂筆記)與報告(如企劃書、論文)的需要，因此本會從「文本活動」及「表達活動」中各挑出了「文本處理」與「報告」兩個題型，作為流利精通級寫作測驗的測驗項目。

本會將此二題型命名為「摘要寫作」與「觀點論述」，前者評量考生能否為長篇文本作出言簡意賅的摘要；後者則檢視考生能否針對複雜的主題撰寫清楚、完整並具有一定篇幅的論說文。「摘要寫作」要求受測者在閱讀一篇 1000 字左右的訪談資料[1]後，再以自己的話語重新組織內容，完成 200 至 300 字的摘要。其評量要點除了重點擷取及組織能力以外，將淺白、口語的訪談文字轉換為高程度書面語的能力亦為檢核要項；「觀點論述」則要求受測者針對爭議性議題，以 800 至 1000 字之篇幅詳述其正反兩面觀點，並提出合宜的因應對策。流利精通級寫作測驗題型分布如表 2 所示。

表 2　測驗題型

題型	題數	字數	時間
摘要寫作	1	200-300	50 分鐘
觀點論述	1	800-1000	120 分鐘

(三)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)。前者根據整體印象，給予一個單一分數，其優點為計分快速，但較為主觀；而後者則針對不同的評量向度，分別給予分數並計算總分，雖費時，但其結果較為客觀，且具信度與效度 (Weigle，2002)。為獲得評分過程的相關證據與提高評分一致性，並掌握教師在各個向度的評分思維，本測驗對具一定篇幅的段落文本均採取分析式評分。以觀點論述題型為例，評分教師依據評分原則和細則，針對「任務完成度」與「語言表現」兩大向度給分，再得出總分。

(四)評分原則

評分原則的制定方法，主要為汲取中外寫作理論相關內容，並參考國際大型外語測驗，如劍橋國際英語認證(Cambridge English)、法語鑑定文憑(DELF-DALF)、歌德德語檢定考試(Goethe-Zertifikat)、德語鑑定測驗(TestDaf)等所制定的寫作評分規則，另外亦諮詢華語文教學與語言測驗相關領域專家學者的

[1] 本會之所以採用書面資料而非語音資料作為試題，乃因一般認為「閱讀輸入」與「寫作輸出」能力之間存在正相關性(如 Grobe & Grobe，1977)，是故採用書面資料，應能降低因試題理解錯誤導致無法正確測出寫作能力的可能性。

意見。流利精通級「摘要寫作」與「觀點論述」兩種題型的評分級距皆設定為 0 至 5 級分；依照兩種題型之文體及評量重點，分別制定出相應的評分原則。

兩種題型的評量向度皆分為「任務完成度」、「語言表現」兩大向度，在任務完成度方面，兩種題型皆將文章的「內容」與「組織」納入評量，然而依寫作要求的不同有所差異：「摘要寫作」的「內容」部分著重從長篇訪談文本中擷取重點的能力，「組織」部分則評核重新架構資訊、銜接文句的技巧等；「觀點論述」的「內容」評量要點則在於舉證說理的適切性及發展性，「組織」則為邏輯性與連貫性。而無論「摘要寫作」或「觀點論述」，在語言表現方面的評分皆包含「句型詞藻表現力」與「詞彙語法正確度」兩大向度，前者檢視能否運用多樣、複雜的句型與高程度詞語，以展現文采；後者則評量能否正確使用詞彙及語法。另外，由於「觀點論述」為篇幅約 800 至 1000 字的長篇文章，在其語言表現向度中，還需評量「段落形式適切度」，即文章是否適當分段。摘要寫作、觀點論述評分原則，如表 3、表 4 所示。

表 3　摘要寫作評分原則

級分	任務完成度	語言表現
5	5.1 訊息近乎完整，組織良好	• 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 • 容許極少數詞彙語法錯誤；容許極少數增字/漏字/錯別字
4	4.1 訊息近乎完整，組織大致良好 4.2 訊息大致完整，組織良好	• 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 • 容許極少數詞彙語法錯誤；容許少數增字/漏字/錯別字
3	3.1 訊息近乎完整，組織良好，但連續數句近乎抄錄原文 3.2 訊息近乎完整，組織良好，但加上少部分作者的觀點 3.3 訊息大致完整，組織大致良好	• 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 • 容許少數詞彙語法錯誤；容許極少數增字/漏字/錯別字
2	2.1 訊息近乎完整，組織大致良好，但連續數句近乎抄錄原文 2.2 訊息近乎完整，組織大致良好，但加上少部分作者的觀點 2.3 訊息大致完整，組織良好，但連續數句近乎抄錄原文 2.4 訊息大致完整，組織良好，但加上少部分作者的觀點 2.5 訊息大致完整，組織大致良好，但連續數句近乎抄錄原文 2.6 訊息大致完整，組織大致良好，但加上少部分作者的觀點	• 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 • 容許少數詞彙語法錯誤；容許少數增字/漏字/錯別字
1	1.1 訊息不足 1.2 組織不佳 1.3 不合規定 (大量抄錄原文；加上過多作者的觀點；以第一人稱(我)撰寫；超過 400 字)	• 句型結構過於簡單；句型變化性不足 • 詞語過於淺白/口語；冗贅重複多 • 詞彙語法錯誤多；增字/漏字/錯別字多 • 標點符號錯誤多，影響閱讀。
0	字數過少(未達 140 字)；完全抄錄原文；完全離題；不知所云；未以中文寫作	

表 4　觀點論述評分原則

級分	任務完成度	語言表現
5	結構嚴謹，銜接策略靈活；舉證與說理適切充實	• 能運用複雜且靈活多變的句型；能大量地運用高程度用語；容許極少數冗贅重複 • 容許極少數詞彙語法錯誤；容許極少數增字/漏字/錯別字 • 分段佳；段落開頭均空兩格(佳)；標點佳
4	結構良好，銜接策略適切；舉證與說理適切充實	• 能運用複雜且多樣的句型；能運用高程度用語；容許極少數冗贅重複 • 容許極少數詞彙語法錯誤；容許少數增字/漏字/錯別字 • 分段佳;1-2 個段落開頭未空兩格(可)；標點佳
3	結構大致良好，銜接策略大致適切；舉證與說理大致適切充實	• 能運用多樣的句型；能運用一般書面用語；容許極少數冗贅重複 • 容許少數詞彙語法錯誤；容許極少數增字/漏字/錯別字 • 分段佳;1-2 個段落開頭未空兩格(可)；標點可
2	結構尚稱良好，銜接策略尚稱適切；舉證與說理尚稱適切充實	• 能運用尚稱多樣的句型；能大致運用一般書面用語；容許少數冗贅重複 • 容許少數詞彙語法錯誤；容許少數增字/漏字/錯別字 • 分段可;1-2 個段落開頭未空兩格(可)；標點可
1	結構不甚理想，銜接策略不甚理想；舉證與說理不甚理想	• 句型結構過於簡單；句型變化性不足 • 高程度詞語/一般書面用語運用不當；詞語過於淺白/口語；冗贅重複多 • 詞彙語法錯誤多;增字/漏字/錯別字多 • 分段不佳；段落開頭均未空兩格;標點符號錯誤多，影響閱讀。
0	字數過少(未達 400 字)；僅抄考題；完全離題；不知所云；未以中文寫作；文體不符（如：寫成應用文、全文對話形式等）；全文條列式（如清單）	

(五)通過門檻

本測驗透過標準設定(standard setting)程序,設定出流利級與精通級之通過門檻。由於給分方式為 0 至 5 級分的多元計分制(polytomous items),與單選題非對即錯的概念不同,通過門檻設定方法乃參考 Yes / No Angoff 法(Impara & Plake, 1997)之概念,並因應測驗形式為建構反應題加以調整。所有標準設定成員皆為華語文及語言學領域專家,程序依循標準化流程執行,各步驟說明如下。

1. 簡介此標準設定之目的與門檻設定的方法。
2. 介紹流利精通級測驗與 CEFR 架構,並說明依據 CEFR 之 C1 及 C2 等級能力描述所定義之流利級與精通級最低能力描述(minimum performance level descriptions)。
3. 說明摘要寫作題型內容與評分原則。
4. 請成員依據提供的流利級、精通級最低能力描述,分別與摘要寫作題型之評分原則進行配對,決定流利級和精通級寫作最低能力表現最為接近評分原則的哪一級分,並寫下判斷依據。
5. 提供成員根據步驟 4 的判斷結果所得之回饋訊息(Cizek & Bunch, 2007)。回饋訊息包含:流利級與精通級 0 至 5 級分的判斷人數,與結果的平均數和標準差。接著,成員們依據上述回饋訊息進行第一回合判斷後討論。
6. 完成第一回合討論後,成員再次以評分原則進行第二回合門檻設定判斷,判斷方式同步驟 4。
7. 根據步驟 6 之第二回合判斷結果,提供成員如步驟 5 之回饋訊息,並進行第二回合判斷後討論。
8. 完成第二回合討論後,成員再次以評分原則進行第三回合門檻設定判斷,判斷方式同步驟 4。
9. 依據成員於步驟 8 所設定之門檻及本測驗發展目的與目標,設定出流利級與精通級摘要寫作題型之通過門檻。

設定流利精通級觀點論述題型之通過門檻時,程序同上述步驟 3 至 9。

完成測驗通過門檻設定後,需檢視標準設定結果是否可靠,即是否具有效度。一般而言,標準設定結果的效度檢核可分為程序性效度、內部效度及外部效

度三部分(Kane，1994)，在此提供程序性效度及內部效度檢核結果。

首先，程序性效度方面，標準設定會議按照既定議程進行，且在各回合間給予與會者充分的分享與討論時間。會議後的問卷(見附件 1)針對評分原則配對的狀況進行調查，共有九題四點量表，平均分數均在 3.8 以上，同意百分比均在 95%以上，此結果顯示與會者均同意自己了解會議目的、會議帶領者對標準設定方法的操作流程及最低能力描述說明清楚、每回合後團體討論和分享，有助於進行下一回合的判斷、對於自己所設定的切截分數有信心等等，可做為程序性效度依據。

內部效度證據則由每一回合通過門檻的標準差作為依據。從表 5 可知，在流利級通過門檻部分，摘要寫作題型的標準差在第一回合兩大向度皆為 0.646，在第二回合則皆為 0.426，第三回合 14 位專家的判斷達到完全一致，標準差為 0；觀點論述則是在第一回合 14 位專家判斷即完全一致，標準差為 0。精通級通過門檻部分，同樣是摘要寫作題型在第一回合標準差較大，兩向度皆為 0.363，而在第二、第三回合達到完全一致，標準差為 0；觀點論述亦在第一回合便已達完全一致，標準差為 0。由上述結果可知，經由判斷後的討論，專家們的意見已趨於一致，也因此觀點論述未進行第二和第三回合判斷。

表 5　標準設定各回合判斷結果之標準差

通過等級	題型	向度	第一回合	第二回合	第三回合
流利級	摘要寫作	任務完成度	0.646	0.426	0.000
		語言表現	0.646	0.426	0.000
	觀點論述	任務完成度	0.000	--	--
		語言表現	0.000	--	--
精通級	摘要寫作	任務完成度	0.363	0.000	0.000
		語言表現	0.363	0.000	0.000
	觀點論述	任務完成度	0.000	--	--
		語言表現	0.000	--	--

華語文寫作測驗流利精通級標準設定結果，在程序性效度與內部效度二項效度證據均獲得支持，即驗證了流利精通級寫作測驗，能有效將華語學習者的寫作表現區分為 CEFR 的 C1 和 C2 兩等級。

流利精通級寫作測驗成績為第一部分摘要寫作及第二部分觀點論述兩題型的分數加總，滿分為 10 分。根據標準設定研究結果，各等級通過分數範圍如表

6 所示。測驗總分介於 4 至 7 分者，可取得流利級(Level 5)證書，總分介於 8 至 10 分者，可取得精通級(Level 6)證書。

表 6　流利精通級通過門檻分數

測驗等級	證書等級	分數範圍
流利精通級	精通級	8-10
	流利級	4-7

三、 測驗標準化流程

測驗的過程必須是客觀化(objective)的，即其結果不應隨施測者或測量情境的不同而改變。欲達到此一目的，就必須嚴訂一套標準化(standardized)的程序，包含測驗編製過程、施測過程、計分與結果的解釋。若測驗的編製者都能依照此流程來進行，對於測驗品質的提升有很大的幫助(陳柏熹，2011)。寫作測驗屬於「表現測驗」(performance assessment)，過去此種測驗常因試題取樣標準不明、評分者的主觀因素、評分流程的客觀因素限制等諸多問題，導致其信度與效度遭受質疑。因此，作為此種高風險測驗(high-stake testing)，必須針對題庫建置與評閱方式，制定「標準化作業流程」(standard operation process；簡稱 SOP)，於測驗內容、程序與評分上皆遵循一套標準化的處理方式。唯所有評分者都能使用同一套標準去評量每一位受測者的能力，並且給予同等公平、公正、客觀的評分，才能確保寫作測驗具有理想的信度與效度。基於此，本測驗建置正式考試製卷流程與評分流程，茲分述如下。

(一) 標準化製卷流程

本測驗正式考試的製卷流程包含：試題的收集、修審、預試、分析、輸入題庫、組合正式卷、檢核正式卷與多媒體檔案七個階段(如圖 1 所示)，茲說明如下。

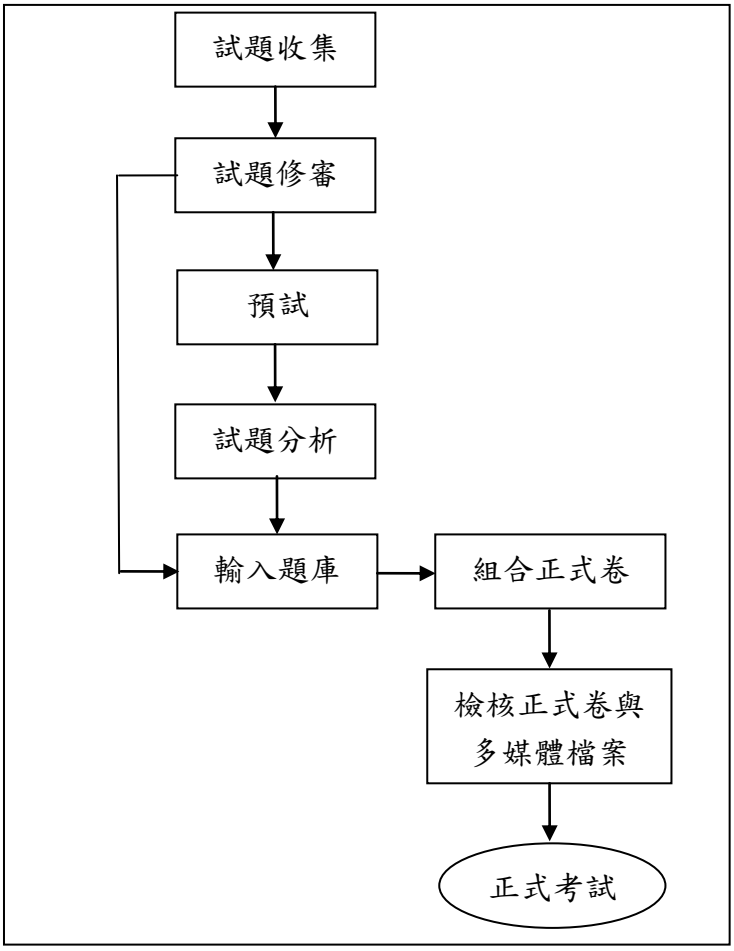

圖 1　正式考試製卷流程

1. 試題收集

本測驗的試題收集工作，主要透過不定期舉辦命題研習，及邀請各華語中心資深教師參與命題兩種途徑來進行。命題前，研發人員提供教師命題相關資料，如：各等級的寫作能力描述、命題方向、題型範例等，作為命題依據。

2. 試題修審

試題修審工作分為三個步驟，首先進行會內初審，而後邀請專家學者外審，最後由研發人員根據外審意見進行修改，並視實際需要製作相關多媒體檔案。各步驟之作業目的簡述如下。

(1) 會內初審

命題教師繳交試題後，由研發人員進行初步審查，其主要目的在於檢視試題是否符合命題原則與相關規定。

(2) 專家學者外審

回收的試題經過研發人員初步修改後，再邀請華語教學及語言測驗相關領域

專家學者進行複審，其審查重點包含：檢視試題所設定的情境與任務之間的邏輯關聯性、個別任務設定的適切性、題意清晰度與流暢性、詞語與語法的正確性等。

(3) 試題修訂與題庫輸入

研發人員根據專家學者的審題意見修訂試題內容，試題確定後，便將題目輸入題庫。

3. 預試

經過命題、修審後的試題進入預試階段，完成樣本收集程序的目的，乃透過量化數據來評估測驗題型是否達到測驗目標，即試題設計是否確能測量受測者實際寫作能力。本會於 2016 年 4 月舉辦流利精通級寫作測驗預試，到考人數為 66 名。

4. 試題分析

經過預試階段的受測者反應將交由本會統計分析人員進行試題分析，並以試題反應理論(Item Response Theory；簡稱 IRT)作為分析取向。由於受測者成績係經由評分教師人工判定，因此受測者成績除了受到其自身具備的寫作能力及試題難度的影響外，還可能受到評分教師評分嚴格度差異的影響。對此，本測驗採用將評分者效果納入估計的多面向模式(facets model) (Linacre，1989)，對考試資料進行分析。由於計分辦法採級分制，屬多元計分方式，因此本測驗使用可進行多面向模式分析的 Facets 3.71.3 版(Linacre，2013)的部分給分模式(partial credit model；簡稱 PCM)對資料進行分析，多面向部分給分模式如公式 1 所示：

$$\log\left(\frac{P_{nijk}}{P_{ni(j-1)k}}\right) = \theta_n - (\delta_i + \tau_{ij} + \eta_k) \quad (1)$$

其中，δ_i 表示第 i 題的整體難度(overall difficulty)；τ_{ij} 表示第 i 題的閾難度(threshold difficulty)或梯級難度(step difficulty)；P_{nijk} 和 $P_{ni(j-1)k}$ 表示第 n 位能力值為 θ 的受測者在第 i 題上被評分者 k 評為 j 分和 $j-1$ 分的機率；η_k 表示評分者 k 的嚴格度，此數值越大表示評分者越嚴格，受測者越難得到高分。

依據 Facets 3.71.3 版輸出報表中的統計指標—訊息加權適配度統計量(inlier-pattern-sensitive fit statistic)之均方(mean-square) (簡稱 Infit MNSQ)，以及偏離反應適配度統計量(outlier-sensitive fit statistic)之均方(mean-square) (簡稱 Outfit

MNSQ)，來評估預試試題品質。因相較於有標準答案的選擇題，寫作測驗的成績還涉及人為評分，影響因素較為複雜，故採取的評估標準為：試題之 Infit MNSQ 與 Outfit MNSQ 數值介於 0.5 至 1.5 者，表示試題適配，亦即試題品質與測驗研發目標一致、試題品質良好。此外，因多面向模式可同時分析測驗中存在的多個面向，以寫作測驗為例，包含評分者嚴格度、試題難度及考生能力三個面向，並可分開呈現估計的結果，故此標準亦可用於評估評分者面向的模式適配情形。

5. 輸入題庫

考量寫作測驗題數較少，若所有試題皆需經由預試階段，較容易有外洩之虞，故本測驗題庫的試題來源分為兩種：一為研發人員依據專家學者審題意見修改的試題；一為經過預試後，顯示試題適配度介於 0.5 至 1.5 之間，且評分較無歧異的試題。經由上述兩種途徑獲得的試題，可確保品質良好，能有效鑑別受測者真實的寫作能力。2016 年度輸入題庫的流利精通級題數為 7 題。

6. 組合正式卷

舉辦考試之前，研發人員自題庫中選取兩種題型的題目，必須涵蓋不同主題，且其題目設定的情境與任務宜避免跟近幾年的考題重覆。

7. 檢核正式卷與多媒體檔案

組卷之後，除研發人員進行試題內容檢核之外，亦進行寫作測驗考試系統測試，以確保考試進行時能夠正常運作。以下說明此階段的工作程序。

(1) 試題內容檢核

組卷後，由研發人員檢核試題的排列順序、格式，以及寫作注意事項的內容。

(2) 電腦考試系統測試

研發人員將檢核無誤的考題製成圖檔，並上傳至考試系統，再登入系統進行模擬交叉測試，檢核內容包含試題的字體大小、間距與版面清晰度等，若有任何問題，立即回報資訊人員進行調整。測試過程分為三個步驟，以下分述各步驟的檢核項目。

I. 登入時：檢查考試流程說明影片的內容是否符合該考試等級。

II. 輸入時：檢查字數統計是否符合題目設定且能正確計算、標點符號列能否正常使用、計時器是否顯示題目設定的時間且能正常運作。

III. 交卷時：考試時間結束或按下「交卷」鈕後，系統是否自動儲存文本。

待上述之檢核項目皆確認無誤後，即完成考試系統測試，製卷流程亦至此結

束，其後將進行正式考試與後續之評分流程。

(二) 標準化評分流程

寫作測驗為主觀性測驗，評分教師需透過完善的培訓過程，方可確保評分的穩定度。本測驗評分教師的養成，分兩階段進行，第一階段為培訓階段，有志成為評分教師的人需先參加本會舉辦的寫作測驗評分研習，以了解本會的評分標準與評閱方式。第二階段為通過評分資格審查階段，即參加過數次評分研習，且確實掌握研習內容的評分教師，才能參與預試或正式考試的評分工作。

上述之評分研習，所邀請之教師來自各大華語中心，具三年以上的華語教學經驗。研習前的籌備工作，主要是從過去測驗的受測者作答反應中，挑選各級分樣卷與提供教師試評的練習卷。研習時，先由研發人員說明評分標準，再請評分教師進行試評與討論。本會從中挑選有熱忱且穩定性高的評分者，做為種子教師，日後邀請其參與正式評閱工作。

預試或正式考試的評閱工作，皆依照標準化流程進行，其流程主要包含評分會議前置作業與舉辦評分會議兩個階段，如圖 2 所示。各階段內容，茲分述如下。

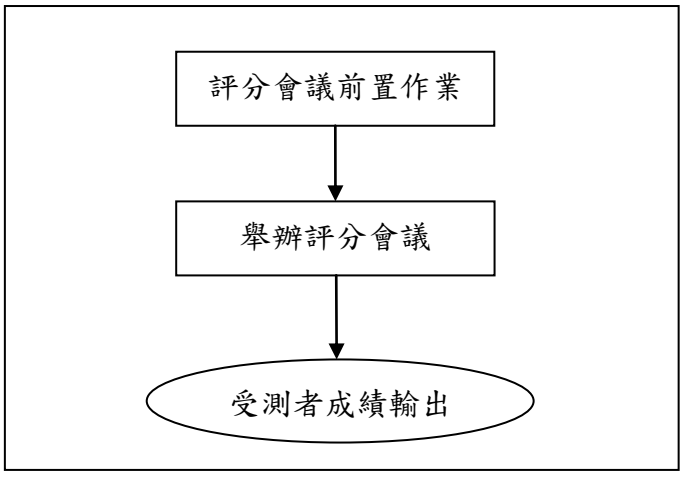

圖 2　評分流程

1. 評分會議前置作業

考試後，研發人員以摘要寫作及觀點論述題型所使用的評分原則為基準，同時輔以該次考試受測者作答反應，草擬寫作任務評分細則，最後搭配評分原則和任務細則挑選各題型各向度級分樣卷並撰寫評語，以供評分教師於正式評閱時作為依據。

2. 舉辦評分會議

評分會議的流程分為兩個步驟，首先說明評分相關規定與試評練習卷，而後進行正式的評分工作。兩步驟之作業流程與目的分述如下。

(1) 說明評分相關規定與試評練習卷

正式評分之前，先由研發人員說明評閱原則、重點與流程，其主要內容包含：各向度的評分標準、偏誤的標註方式、各級分樣卷的說明，以及評分系統的操作方式。而後請評分教師依據上述內容進行試評，透過試評與討論，協助評分教師切實掌握評分要領，並調整各自的評分寬嚴度，以利提高評分一致性。

(2) 正式評分

完成上述程序，始可進行正式評閱工作。每一份考生作答反應至少分派予二位評分教師。評分會議結束後，由研發人員彙整所有評分，並針對分數差距較大者進行討論，以決定最後的成績。

本會建置了「寫作測驗線上評分系統」以便教師進行評分。為避免評分標準偏離，研發人員可由系統後台即時檢視各評分教師的評閱情形，若發現評分問題，可立即向該教師反應、討論，待釐清後再繼續評閱。教師之評閱結果直接儲存於線上系統當中，研發人員可於閱卷結束後，匯出評閱結果進行後續分析。

四、 測驗評估

一份測驗是否能夠發揮效用，並能確切地測量受測者的目標潛在能力，通常可通過該測驗的信度與效度分析來進行整體性評估。緣此，本節將討論 2016 年 11 月 6 日所舉辦的流利精通級正式考試之信效度來說明寫作測驗之效能。

本次考試共有 57 名考生到考，第一部分摘要寫作題型有六位評分者參與評分工作，所有評分者皆評閱 57 篇；第二部分觀點論述題型則由七位評分者進行評分，評閱資料數介於 28 至 57 篇不等。由於本測驗採用多面向模式的主要目的為評估評分者的評分一致性以及試題整體難度，故以下報告只針對評分嚴格度與試題難度做討論。此次測驗信、效度分析結果將詳述如下。

(一) 信度

所謂信度，指的是測驗結果(即成績)的穩定性與一致性。一份測驗，倘若無論何時、何地，由任何人進行施測、計分，均能得到一致性很高的測驗結果，則表示此份測驗具有較高的信度。換言之，該測驗所獲得之測驗結果測量誤差很小(或稱精準性高)。

一般而言，常被用來評估測驗信度的指標主要有「再測信度」、「複本信度」、「內部一致性信度」、「評分者信度」四類。其中，再測信度主要觀察在不同時間點施測時，所獲得的測驗成績是否具有一致性；複本信度用來觀察以不同題本施測，所獲得之測驗成績是否具有穩定性；內部一致性信度主要觀察測驗所測量之潛在特質是否具有一致性；評分者信度則是關注經由不同評分者所得到的評分結果是否具有一致性。

本測驗屬於建構反應型能力測驗，受測者成績之取得，主要仰賴評分者實際進行人工評分，評分辦法相對主觀。緣此，評分者的「評分一致性」遂為影響受測者分數之主要因素。「評分者的評分一致性」大致可分為「評分者間一致性」(inter-rater consistency)及「評分者內一致性」(intra-rater consistency)兩種類型。前者指的是不同評分者在評量相同受測者時，其評量分數(或分數等級)的一致性；後者則是指同一評分者在評量給分上的一致性(或穩定性)。

本節主要說明 2016 年流利精通級寫作測驗正式考試之信度，將以「評分者嚴格度變異」和「斯皮爾曼等級相關」來評估評分者間信度與評分者內信度。

1. 評分者間信度

(1) 嚴格度變異分析

此部分採用 Facets 3.71.3 版的多面向部分給分模式對資料進行分析，檢視評分者嚴格度差異，以下將分別說明第一部分與第二部分之評分者嚴格度結果。

第一部分摘要寫作題型的評分者嚴格度如表 7 所示，評分者 C08 給分較為嚴格，嚴格度為 0.415；而評分者 C01、C02 給分略為寬鬆，嚴格度均為-0.200。以嚴格度平均值 0 作為標準來看，六位評分者嚴格度均介於±0.5 logit 以內，表示六位評分者嚴格度相當一致。

表 7　第一部分摘要寫作題型評分者嚴格度

評分者編號	篇數	觀察的平均值	嚴格度	標準誤 (S.E.)	Infit MNSQ	Outfit MNSQ
C08	57	2.14	0.415	0.122	0.85	0.75
C07	57	2.29	0.173	0.117	0.70	0.59
C03	57	2.44	-0.050	0.113	1.09	1.09
C04	57	2.50	-0.138	0.112	1.15	1.06
C01	57	2.54	-0.200	0.111	0.93	0.85
C02	57	2.54	-0.200	0.111	1.27	1.47

第二部分觀點論述題型的評分者嚴格度如表 8 所示，七位評分者中，評分者 C07 給分較為嚴格，嚴格度為 0.771，而評分者 C02 給分較為寬鬆，嚴格度為-0.648 以嚴格度平均值 0 作為標準來看，七位評分者中，有五位嚴格度介於±0.5 logit 以內，評分標準較為一致；而 C07 評分較為嚴格，C02 評分較為寬鬆。

表 8　第二部分觀點論述題型評分者嚴格度

評分者編號	篇數	觀察的平均值	嚴格度	標準誤 (S.E.)	Infit MNSQ	Outfit MNSQ
C07	30	1.55	0.771	0.171	0.79	0.57
C03	57	1.73	0.366	0.115	0.97	0.92
C01	28	2.00	0.087	0.157	0.50	0.50
C08	57	2.08	-0.122	0.107	0.71	0.64
C04	56	2.08	-0.150	0.108	0.85	0.81
C11	46	2.27	-0.304	0.116	1.45	1.67
C02	35	2.53	-0.648	0.134	1.93	1.91

(2) 斯皮爾曼等級相關

針對第一部分摘要寫作與第二部分觀點論述各評分者之給分，進行斯皮爾曼等級相關分析以了解兩兩評分者之間的信度。由於流利精通級測驗採分析式評分，故除了整體級分外，亦呈現評分向度的相關分析結果。結果如表 9 所示，因篇幅有限，僅呈現平均數、標準差等描述性統計結果。

摘要寫作題型方面，各個評分者整體級分之相關係數介於.69 至.87 之間，相關係數平均數為.78，任務完成度與語言表現兩向度的平均數分別為.70 與.61。觀點論述題型方面，評分者兩兩之間整體級分的相關係數介於.49 至.87 之間，平均數為.71，任務完成度和語言表現的相關係數平均數分別為.71 和.70。

一般來說，相關係數達.40 以上即表示具有中度相關，達.70 以上表示有高度相關。顯示流利精通級寫作測驗評分者間信度大致良好，摘要寫作題型的語言表現向度相關係數平均數較低，未來將針對此向度加強培訓，以讓評分者更為確實掌握評分規準。

表 9　評分者間斯皮爾曼等級相關

題型	評分向度	平均數	標準差	最小值	最大值
摘要寫作	整體級分	0.78	0.06	0.69	0.87
	任務完成度	0.70	0.07	0.60	0.89
	語言表現	0.61	0.13	0.39	0.80
觀點論述	整體級分	0.71	0.12	0.49	0.87
	任務完成度	0.71	0.08	0.51	0.84
	語言表現	0.70	0.13	0.43	0.87

2. 評分者內信度

此部分透過嚴格度分析中評分者之標準誤以及適配性指標（INFIT MNSQ）數值檢視評分者本身給分的一致性情形。由表 7、表 8 評分者嚴格度標準誤所提供的直接證據顯示，摘要寫作題型各評分者之標準誤介於 0.111 至 0.122 之間；觀點論述題型各評分者之標準誤介於 0.107 至 0.171 之間。整體而言，兩題型六名和七名評分者的標準誤變異情形差異不大，表示評分者皆具有自身評分的穩定性。

而由 Infit MNSQ 及 Outfit MNSQ 評估評分者自身給分一致性之間接證據也顯示，摘要寫作題型六名評分教師均符合評估標準，數值介於 0.5 至 1.5 之間；

觀點論述題型七名評分教師中，有六名數值介於 0.5 至 1.5 之間，顯示這些評分者內一致性佳，給分符合模式預期，評分穩定性良好。至於評分者 C02 之 Infit MNSQ 與 Outfit MNSQ 均大於 1.5，顯示其自身評分標準較不一致。

由上述評分者信度分析結果可知，2016年流利精通級寫作測驗正式考試的評分者大體上具有評分者內一致性，大多數評分者自身給分穩定度良好；評分嚴格度方面，第一部分題目與第二部分題目大部分評分者的評分者嚴格度均介於±0.5 logit以內，比例分別為100%與71.4%；斯皮爾曼等級相關結果，第一部分題目與第二部分題目兩兩評分者間相關係數的平均值達中度或高度正相關，評分者間信度大致良好。

為了確保評分教師的評分品質，針對評分結果較不理想，如偏嚴格、偏寬鬆或與最終評定成績較不一致之評分教師，將列入觀察名單並再給予訓練，若後續評分狀況仍未改善，即不續聘。

(二) 效度

所謂測驗效度，指的是檢驗一項測驗是否能測量到欲測量的能力(或潛在特質)。由於目標測量能力無法被直接觀察，因此，測驗效度皆須藉由受測者在試題上的作答反應或行為來間接推估。通常用來驗證測驗效度的證據主要分為「內容效度」(content validity)、「建構效度」(construct validity)、「效標效度」(criterion validity)三大類。其中，內容效度指的是測驗內容的相關證據；建構效度為關於測驗架構的證據；效標效度則是指測驗結果預測力的相關證據。

本測驗是一種「表現測驗」，受測者的成績由評分者依據評分原則進行判定，評分者的主觀判斷即為評分之主要影響因素。也就是說，若評分者不能確實掌握評分原則來進行評分，則將無法正確區分受測者能力，並連帶影響測驗效度。因此，在寫作測驗中，讓評分者接受一系列標準化程序的評分訓練，是相當重要的一環。此一標準化程序被稱為程序性效度(procedural validity)，可確保測驗相關內容皆是經由標準化程序而來，以作為內容效度的證據。通過測驗試題分析、探索性因素分析，研究人員可評估測驗試題所測量到的能力是否與測驗發展時所定義的架構或內容相吻合，此屬建構效度的證據。

以下將分別以程序性效度、試題分析、探索性因素分析結果來描述本次正式考試的內容效度及建構效度。

1. 程序性效度

本會制訂的寫作測驗評分標準化流程，分兩個階段進行。第一階段為評分會議前置作業，研發人員先根據試題設定的「寫作任務」草擬評分細則，並邀請資深評分教師進行試評，再參考試評意見加以修改，最後依據修訂後的評分細則挑選各級分樣卷、標準卷及練習卷；第二階段是舉辦評分會議，在正式評分之前，先由研發人員說明評分標準，再請評分教師進行試評與討論，建立共識後，才進行正式評分。在評分過程中，研發人員透過評分系統後臺監控評分狀況，必要時，即時提供評分回饋，以利評分教師調整其嚴格度。

評分會議結束後，由統計人員進行評分結果分析，提供評分嚴格度、評分者間與評分者內一致性等分析資料，作為未來評分訓練之參考。標準化評分會議的工作內容，參見表 10。

表 10　標準化評分會議的工作內容

階段	工作項目	內容
一	評分會議前置作業	1. 草擬任務評分細則、試評與修改。 2. 挑選各級分樣卷、標準卷與練習卷。
二	舉辦評分會議	1. 說明評分相關規定、試評、討論。 2. 正式評分，並提供評分回饋。

寫作測驗按照標準化評分流程進行評分，評分教師的評分嚴格度以平均值 0 作為標準來看，第一部分題目六位評分者嚴格度均介於±0.5 logit 以內；在第二部分題目的七位評分教師當中，有五位評分者嚴格度介於±0.5 logit 以內，一位偏嚴，一位偏鬆。顯示大多數的評分者嚴格度較為接近，且所有評分教師皆具有自身評分一致性，也就是說，各評分教師在評分上具有穩定度。由此可知，標準化評分程序可有效訓練評分教師依據評分準則進行評分，從而達到評分之一致性。

2. 試題分析

本測驗之試題分析方式是採試題反應理論(IRT)。試題反應理論的一項重要假設為：單向度假設。所謂單向度假設，指的是測驗中所有題目皆在測量相同潛在特質，當受測者回答試題並非仰賴單一特質時，若忽略此一訊息並進行單向度試題反應理論分析，所獲得的試題參數及受測者能力估計值將是具有偏誤的。

本節將採用 Facets 3.71.3 版的多面向部分給分模式(如公式 1 所示)對資料進行分析，結果如表 11、表 12 所示，第一部分摘要寫作題型和第二部分觀點論述題型的兩個評分向度，任務完成度皆略難於語言表現，難度分別為 0.233 和 -0.233，以及 0.201 和-0.201。此結果可能由於相較於任務完成度來說，語言表現為寫作能力的基礎，故難度略低。

採用 Infit MNSQ 及 Outfit MNSQ 介於 0.5 到 1.5 的標準評估試題是否與單向度試題反應理論模式適配，結果顯示兩個題型的任務完成度、語言表現兩個評分向度數值均符合標準，與模式的適配情形均為良好，顯示不同評分向度測量到相同的潛在特質，也就是寫作表達能力。綜上所述，本測驗流利精通級正式考試具有一定程度的建構效度。

表 11　第一部分評分向度難度分布

向度	難度	標準誤(S.E.)	Infit MNSQ	Outfit MNSQ
任務完成度	0.233	0.071	1.06	1.04
語言表現	-0.233	0.062	0.97	0.89

表 12　第二部分評分向度難度分布

向度	難度	標準誤(S.E.)	Infit MNSQ	Outfit MNSQ
任務完成度	0.201	0.067	0.84	0.79
語言表現	-0.201	0.066	1.23	1.21

3. 探索性因素分析

由於流利精通級寫作測驗的題型分為兩部分，較適合以探索性因素分析評估測驗的建構效度，故採用 SPSS (23.0 版本)以摘要寫作和觀點論述兩題型得分作為測量變項進行分析。首先進行 KMO (Kaiser-Meyer-Olkin)與 Bartlett’s 球形檢定，以確認資料是否適合進行因素分析。流利精通級寫作測驗 KMO 數值為 0.500，符合 Kaiser (1974)所建議至少需達 0.5 之標準。Bartlett’s 球形檢定結果，數值為 50.131，達顯著水準(p<.001)。此兩項結果皆表示資料的相關矩陣存在共同因素，適合進行因素分析。

圖 3 為因素分析陡坡圖，由下圖可知於流利精通級寫作測驗抽取出一個因素後，線條即大幅降低；因素分析結果，抽取出因素特徵值大於 1.0 者(特徵值 1.776)，該因素解釋變異量達 88.8%。由於摘要寫作與觀點論述兩題型皆為測量

寫作表達能力，因此將該因素命名為「寫作表達能力」，而兩題型的因素負荷量皆為.94，上述分析結果提供了流利精通級寫作測驗的建構效度證據。

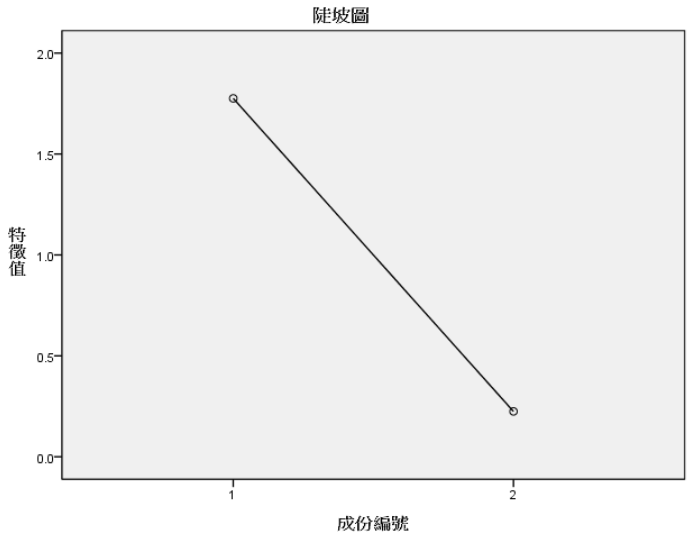

圖 3　流利精通級寫作測驗因素分析陡坡圖

五、結論

本測驗 2016 年技術報告，首先簡介流利精通級寫作測驗的研發相關內容，如能力描述、測驗題型、評分方式、評分原則與通過門檻等，其次說明標準化的製卷與評分作業流程，最後分析流利精通級寫作測驗正式考試的信效度，並根據各項分析結果提出相關討論及建議。

在測驗信度方面，本會透過「評分者嚴格度變異」及「斯皮爾曼等級相關」來進行評分者間信度分析與評估評分者內一致性；在測驗效度方面，為使受測者獲得符合其寫作能力之分數，本會制定標準化的製卷與評分作業流程，藉此確保測驗相關內容皆經由標準化程序而來，而此程序效度即為本測驗提供內容效度方面的證據。在此應補充說明的是，由於本測驗的受測者成績主要仰賴評分教師判定，因此，受測者成績除了受到受測者自身具備之寫作能力與測驗試題難度的影響之外，同時也受到評分教師嚴格度變異的影響，因此評分教師自身給分穩定性與評分教師間給分一致性，對於受測者成績而言便相當重要。為了確保評分教師確實掌握寫作測驗的評分標準，給予受測者適切的評分，本會針對評分較嚴格或與本會最終評定成績較不一致的評分教師，進行進一步的培訓。若評分狀況仍未見改善，將列入觀察名單或不予續聘。為使評分教師了解其自身評分狀況，本會在評閱工作結束後，皆提供評分教師評分嚴格度等相關回饋。

除了具備測驗內容效度方面的證據之外，在施測完成後，本會統計分析人員亦根據受測者作答反應資料進行試題分析，其主要目的在於檢核受測者之反應資料所建構出的測驗架構，是否與本測驗制訂的研發目標相同，並以此作為測驗之建構效度證據。

由 2016 年度全國性流利精通級寫作測驗正式考試的信度與效度分析資料來看，可大致總結以下兩項要點：

(一) 建置標準化的評分作業流程，有助於提高評分者自身評分穩定性及評分者間評分一致性。

(二) 受測者獲得的測驗成績與本測驗所訂定的目標寫作能力相符。

綜上所述，2016 年度流利精通級寫作測驗可測得受測者之目標寫作能力，故受測者成績具有可信度。

六、文獻

陳柏熹(2011)。**心理與教育測驗：測驗編製理論與實務**。台北：精策教育。

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**(編號：ISBN 978-986-92167-5-3)。新北市：國家華語測驗推動工作委員會。

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**(編號：ISBN 978-986-93763-0-3)。新北市：國家華語測驗推動工作委員會。

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

Council of Europe. (2001). *Common European Framework of Reference for Languages: Learning, Teaching, Assessment*. Cambridge: Cambridge UniversityPress.

Grobe, S. F., & Grobe, C. H. (1977). Reading skills as a correlate of writing ability in college freshmen. *Reading World*, 16, 50-54.

Impara, J. C.,& Plake, B. S. (1997).Standard-setting: An alternative approach. *Journal of Educational Measurement,34*,353–366.

Kane, M.(1994). Validating the performance standards associated with passing scores. *Review of Educational Research, 64*, 425–461.

Kaiser, H. F. (1974). An index of factorial simplicity. *Psychometrika, 39*, 31-36.

Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago：MESA.

Linacre, J. M. (2013). Facets (Version 3.71.3) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Weigle, S. C. (2002). *Assessing writing.* Cambridge: Cambridge University Press.

附件 1　流利精通級寫作測驗標準設定研究問卷調查結果

	問卷內容	平均數	同意百分比
評分原則配對	1. 我了解本次 TOCFL 寫作測驗與 CEFR 對應研究會議的目的。	3.9	98%
	2. 會議帶領者對於標準設定方法的流程說明得很清楚。	4.0	100%
	3. 會議帶領者對於本研究 Angoff 法的進行方式說明得很清楚。	3.9	98%
	4. 會議帶領者對於 CEFR C1 與 C2 最低能力描述說明得很清楚。	4.0	100%
	5. 第一回合後團體中的討論和分享，有助於我進行第二回合的判斷。	3.9	98%
	6. 第二回合後團體中的討論和分享，有助於我進行第三回合的判斷。	3.8	95%
	7. 我確實根據 C1 最低能力描述判斷評分原則 C1 的通過門檻分數。	3.8	95%
	8. 我確實根據 C2 最低能力描述判斷評分原則 C2 的通過門檻分數。	3.9	98%
	9. 整體來說，我對於自己所設定的通過門檻分數(cut score)有信心。	3.8	95%

註：問卷填答方式，1 表示非常不同意；2 表示不同意；3 表示同意；4 表示非常同意。

書　　名：華語文能力測驗技術報告—2016 (2)
寫作測驗信效度
出 版 者：國家華語測驗推動工作委員會
24449 新北市林口區仁愛路一段 2 號
886-2-7734-5638
印 刷 者：上校文化印刷有限公司
80744 高雄市三民區通化街 88 巷 26 號
886-7-311-6011
出版日期：2018 年 4 月
定　　價：新台幣 100 元