華語文能力測驗技術報告 ─2013(4)

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佈之標準化流程、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)是由「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)進行研發，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。本測驗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。本會已於 2011 年 10 月，在臺灣地區推出基礎級與進階級正式考試。

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[1]，而每一等級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。此一架構相較於僅能區分受測者是否通過測驗而言，能更進一步分辨出通過測驗者群其能力的高低；同時，對於應試者及試務工作者來說，更符合經濟效益，並能提高測驗效能。例如：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，也就是一份測驗可同時判斷兩個等級程度。

本報告包含三個部分，首先簡介 2013 年進階高階級寫作測驗正式考試的能力描述、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 2013 年進階高階級正式考試資料進行分析，並評估此一年度的寫作測驗整體性信度與效度。

[1]入門基礎級、流利精通級測驗尚在研發中，預計 2014 年推出入門基礎級正式考試。

二、 簡介

2013 年度華語文寫作測驗正式考試等級為進階高階級(Band B)，依照測驗成績可區分為進階級(Level 3)與高階級(Level 4)兩個等級，分別對應 CEFR 的 B1(Threshold)和 B2(Vantage)。該測驗等級之能力描述、測驗題型、評分方式、評分原則與通過門檻，茲分述如下。

（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，在寫作表達能力方面，B1 等級學習者能對自己感興趣的熟悉主題，以連結一連串零散短句的方式，撰寫淺白且連貫的文章；B2 等級學習者則是能對各種自己感興趣的主題，撰寫清楚且詳盡的文章，並能統整與評估不同來源的訊息和論點。而在寫作互動能力上，B1 等級學習者能書寫私人信件和便條，以要求或傳達與切身相關的簡單訊息，並使他人理解自己想要表達的重點；能對具體和抽象的主題，相當精確地傳遞訊息和想法，也能核對訊息的正確性，並針對問題提出詢問或解釋。至於 B2 等級學習者則是能有效地傳遞消息或觀點，並與他人的看法作連結。

綜合比較以上兩個等級的寫作能力描述，可推知 B1 等級學習者能傳達較為熟悉和常見的任務內容，尚不具備推理和論證能力，而 B2 等級學習者則是能就假設的情況進行論證。本測驗依此制定進階級與高階級寫作能力描述，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
進階級	能寫出較為詳細的私人信件，藉由描述經驗、情感、事件等，傳達切身相關的訊息。
高階級	能撰寫闡述論點的文章或報告，對特定觀點提出支持或反對的理由，並解釋各種面向的優劣。

（二）測驗題型

進階高階級寫作測驗題型是參考 CEFR 的 B1 與 B2 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，分別是表達活動、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」；文本形式包括：「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不同。

為了在有限的測驗時間內可有效測出並區別 B1、B2 兩個等級受測者之寫作能力，研發人員根據 CEFR 對 B1 及 B2 所訂的寫作能力描述內容，評估了各類寫作活動之難度適切性，並據此制定出兩種題型。在上述三類寫作活動中，「便條」、「留言」及「表格」此類題型，通常作答內容較為簡短，所需的寫作能力層次較低，偏向入門基礎級水準；而「筆記」及「文本處理」則主要為要求受測者從說話者談話內容或文本內容記下重點、摘要，或是改寫文章，此類題型需要統整不同來源之文本，其認知複雜度相對較高，屬於流利精通級水準。以上述及之寫作活動皆較不適合用以評量進階高階級寫作能力。

至於「書信」形式，在互動類別中，B1 書信寫作能力為「能書寫較為詳細的私人信件，以描述經驗、情感及事件」，此外，書信寫作能力也涵蓋其他寫作溝通類別測驗重點，如：「創作」中提到的能力「能以簡單、連貫的文章，描述經驗、情感及反應」；「報告」及「論文」提到「能傳達例行事項的真實訊息及陳述行動的理由」；「便條」、「留言」及「表格」中提到「傳達與切身相關的簡單訊息，並使他人理解自己想要表達的重點」。因此，第一大題選擇以「書信寫作」作為評量受測者段落寫作能力之測驗題型，要求完成一封 250 至 350 個字的私人信件，以檢視受測者是否能針對與切身相關的主題，以淺白且連貫的敘述方式，傳遞一般訊息及個人經驗。

第二大題根據 CEFR 對 B2 寫作能力之描述，能對自己感興趣的主題，撰寫清楚且詳細的文章，針對特定觀點提出支持或反對的理由。將此寫作能力描述對應到 CEFR 寫作活動溝通類別則屬「報告」類之題型。是故，選擇以「觀點論述」作為評量受測者撰寫論點寫作能力之測驗題型，要求須完成 500 至 600 個字的論述性文章，以檢視受測者是否能針對與自身相關的主題，提出自己的觀點，並加

以解釋。進階高階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
書信寫作	1	250-350	40 分鐘
觀點論述	1	500-600	60 分鐘

（三）評分方式

寫作測驗的評分方式，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)。本測驗為獲得評分過程的相關證據及掌握教師在各個向度的評分思維，以提高評分一致性，而採取分析式評分。以書信寫作題型為例，評分教師依據評分原則和任務細則，分別針對寫作任務、結構組織句法表現、詞語表現三大向度，給予受測者適當的分數，最後再計算出總分。

（四）評分原則

本測驗評分原則的制定方法，主要汲取中外寫作理論相關內容，以及參考國際大型外語測驗，如劍橋英檢、法語檢定、德語檢定等，所制定的寫作評分規準，並諮詢華語文教學與語言測驗相關領域專家學者的意見。進階高階級將書信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不同，制定出兩套評分原則。此兩種題型在「結構組織句法表現」及「詞語表現」都有其相對應的評分規範，但依照文體的不同，在任務完成度上的側重點也不同。書信寫作的評量重點在於受測者訊息的傳遞是否完整清楚；而觀點論述的評量重點在於論述的立場是否前後一致、脈絡是否清楚、論證是否具說服力。各題型的評量向度係參考文體特點和受測者真實文本特徵進行劃分，書信寫作題型的評量向度分為「任務完成度」、「結構組織句法表現」和「詞語表現」三大向度。其中，任務完成度主要檢視受測者是否依據題意選取適切素材加以發

展，以完成題目設定的溝通任務；結構組織句法表現主要檢視受測者的文章形式概念、文句銜接能力以及句內結構的掌握度；詞語表現主要檢視受測者對詞語的掌握程度。觀點論述題型則是分為「任務完成度」與「形式概念語言能力」兩大向度。其中，任務完成度主要檢視受測者於表達個人觀點時，其立場是否具一貫性；組織邏輯是否緊密清楚；所提出的論點與論述能否表達立場且有所發展；舉證是否切合論點且具說服力。書信寫作與觀點論述兩種題型的評分原則，分別如表 3 和表 4 所示。

表 3　書信寫作題評分原則

級分	任務完成度	結構組織句法表現	詞語表現
5	● 詳細回應所有任務(依任務要求選取適切素材加以發展，有效達成溝通目的)。	● 形式佳(開場與結尾適切；結構佳；分段適切；段落間或任務間銜接良好；標點符號使用大致正確)。 ● 文句銜接良好 ● 句內結構錯誤極少	● 詞語佳(誤用/自創/錯別字/增字/漏字極少)。 ● 冗詞贅句極少。
4	● 大致回應所有任務。 ● 詳細回應大部分任務，少部分任務過於簡略。	● 形式大致良好(開場與結尾適切；稱謂未頂格或內文開頭未空兩格；分段大致適切；段落間或任務間銜接大致良好；標點符號使用尚可)。 ● 文句銜接大致良好。 ● 少數句內結構錯誤。	● 詞語大致正確。 ● 偶有語意重複或贅述。
3	● 簡單回應所有任務。 ● 詳細回應大部分任務，然未回應其中一項。 ● 大致回應大部分任務，然少部分任務過於簡略。	● 形式尚可(開場或結尾缺/不適切；缺稱謂或署名；分段較多；段落間或任務間銜接尚可；標點符號使用不甚理想)。 ● 文句銜接尚可。 ● 句內結構尚可。	● 詞語尚可(有些錯誤，但不影響理解)。 ● 語意重複或贅述略多。
2	● 所有任務發展不足。 ● 大致回應大部分任務，然其中一項未回應。 ● 簡單回應大部分任務，然少部分任務過於簡略。	● 形式不佳(開場與結尾缺/不適切；稱謂或署名與內文在同一段落；分段過多或完全未分段；段落間或任務間銜接不佳；標點符號使用差，錯誤影響理解)。 ● 文句銜接不佳；通篇使用基礎銜接詞語。 ● 句內結構不佳。	● 詞語差(錯誤影響理解)。 ● 語意重複或贅述較嚴重。
1	● 所有任務內容貧乏。 ● 未回應大部分任務。	● 形式極差(僅有開頭；稱謂與署名缺/錯置/非題目設定的姓名；段落間或任務間銜接極差；標點符號使用極差，錯誤嚴重影響理解)。 ● 脈絡凌亂，難以理解。 ● 句內結構極差。	● 詞語極差(錯誤嚴重妨礙理解)。 ● 語意重複或贅述嚴重。
0	完全空白；僅抄題目；文不對題；全文對話形式；全文條列式；字數極少。		

表 4　觀點論述題評分原則

級分	任務完成度	形式概念語言能力
5	● 立場鮮明[2]，脈絡佳[3]，論證佳[4]。	● 形式佳(首尾俱全，段落開頭空兩格；分段適切；標點符號使用適切)。 ● 句內結構錯誤極少。 ● 詞語誤用極少，偶見文白夾雜。 ● 句型、詞語靈活多樣。
4	● 立場鮮明，脈絡佳，論證一般。 ● 立場鮮明，脈絡一般，論證佳。 ● 立場清楚[5]，脈絡佳，論證佳。	● 形式大致良好(首尾俱全，段落開頭未空兩格；段落大致分明；標點符號使用大致適切)。 ● 句內結構錯誤少。 ● 詞語大致正確，文白夾雜較多。 ● 句型、詞語變化較多。
3	● 立場鮮明，脈絡一般，論證一般。 ● 立場清楚，脈絡一般，論證佳。 ● 立場清楚，脈絡佳，論證一般。 ● 立場清楚，脈絡一般，論證一般。	● 形式尚可(分段、標點尚可)。 ● 句內結構尚可。 ● 詞語尚可。 ● 句型、詞語變化尚可。
2	● 立場不夠清楚。 ● 立場清楚，但無引論或結論。 ● 立場清楚，但脈絡不佳。 ● 立場清楚，但論證不佳。	● 形式不佳(文末缺句點或幾個字，但文意完整；分段較多；標點符號使用不佳)。 ● 句內結構較差。 ● 詞語較差。 ● 句型、詞語變化少；過於口語。
1	● 無立場或前後矛盾。 ● 引論與結論皆無。 ● 大部分內容不知所云。 ● 論證不清或薄弱。	● 形式極差(僅有開頭；非篇章形式、分段過多或未分段；標點極差)。 ● 句內結構極差。 ● 詞語極差。 ● 句型、詞語單調重複或多屬基礎級。
0	完全空白；僅抄題目；文不對題；文體不符；全文對話形式；全文條列式；字數極少；不知所云。	

[2]「立場鮮明」指引論、結論前後一致，且結尾處理得當，能突顯一貫立場。

[3]「脈絡佳」指組織緊密，邏輯清楚，銜接策略佳。

[4]「論證佳」指論點能充分表明立場，且論述充實並有發展，兼具廣度與深度，同時舉證佳，不但切合論點，且說服力強，足以支持論述。

[5]「立場清楚」指有一貫立場，但引論或結論不佳。

(五) 通過門檻

本測驗透過標準設定(standard setting)程序，設定出進階級與高階級之通過門檻。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 法之概念，再因應測驗形式為建構反應題加以調整。所有標準設定成員均由華語文及語言學領域專家所組成，並依循標準化流程執行，標準設定程序各步驟說明如下。

1. 簡介此標準設定之目的與門檻設定的方法。
2. 介紹進階高階級測驗與 CEFR 架構，並說明依據 CEFR 之 B1 及 B2 等級能力描述所定義之進階級與高階級最低能力描述(minimum performance level descriptions)。
3. 說明書信寫作題型內容與評分原則。
4. 請成員依據提供的進階級、高階級最低能力描述，分別與書信寫作題型之評分原則進行配對，決定進階級和高階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. 請成員閱讀書信寫作題型的 10 篇受測者文本後，依據進階級、高階級最低能力描述，分別判斷每篇文本的 CEFR 等級(B2、B1、不到 B1)，並寫下判斷依據。
6. 提供成員根據步驟 4 及 5 的判斷結果所得之回饋訊息(Cizek& Bunch，2007)。回饋訊息包含：(1)進階級與高階級 0 至 5 級分的判斷人數，與結果的平均數和標準差；(2)每篇文本被判定為 CEFRB2、B1、不到 B1 等級的人數。接著，成員們依據上述回饋訊息進行第一回合判斷後討論。
7. 完成第一回合討論後，成員再次以評分原則和文本進行第二回合門檻設定判斷，判斷方式同步驟 4 及 5。
8. 根據步驟 7 之第二回合判斷結果，提供成員如步驟 6 之回饋訊息，並進行第二回合判斷後討論。
9. 完成第二回合討論後，成員再次以評分原則和文本進行第三回合門檻設定判斷，判斷方式同步驟 4 及 5。
10. 依據成員於步驟 9 所設定之門檻及本測驗發展目的與目標，設定出進階級與高階級書信寫作題型之通過門檻。

設定進階高階級觀點論述題之通過門檻時，程序同上述步驟 3 至 10。

完成測驗通過門檻設定後，需檢視標準設定結果是否可靠，具有效度。一般來說，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。會議後的問卷調查結果顯示(見附件 1)，與會者均同意會議帶領者對會議目的/任務解釋清楚、對標準設定方法的操作流程說明得很清楚、能了解最低能力者在標準設定方法的涵義、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：1.每一回合通過門檻的標準差；2.每一回合文本 CEFR 等級判斷與實際級分之斯皮爾曼等級相關作為依據。標準差部分，從表 5 可知，在進階級通過門檻部分，書信寫作和觀點論述兩個題型的標準差在第一回合較大，在第二回合 11 位專家的判斷已達到完全一致，標準差為 0；高階級通過門檻部分，書信寫作題在第一回合標準差較大，第二和第三回合標準差微幅降低，惟觀點論述題，經過第一回合討論後，第二、第三回合均有專家調降門檻，但因有兩位專家維持第一回合意見，故第二、第三回合的標準差反而較第一回合略高，相差 0.009。即使如此，整體上經由判斷後的討論，專家們的意見越趨一致。

表 5　標準設定各回合判斷結果之標準差

通過等級	題型	第一回合	第二回合	第三回合
進階級	書信寫作	0.302	0.000	0.000
	觀點論述	0.221	0.000	0.000
高階級	書信寫作	0.405	0.316	0.316
	觀點論述	0.396	0.405	0.405

書信寫作與觀點論述兩個題型各 10 份文本 CEFR 等級判斷，與實際級分的斯皮爾曼等級相關分析，將不到 B1、B1、B2 分別編碼為 0、1、2，與文本實際級分求相關的結果，書信寫作題三個回合的相關係數依序為.80 至.95 (p<.01)、.82 至.95 (p<.01)、.80 至.95 (p<.01)，觀點論述題三個回合的相關係數依序為.60 至.88 (p<.01)、.72 至.88 (p<.01)、.72 至.88 (p<.01)，顯示專家們對於文本通過等級的判

斷與實際得分之間具有中度或高度的正相關存在，判斷結果與實際得分頗為一致。

華語文寫作測驗進階高階級標準設定結果，在程序性效度與內部效度二項效度證據均獲得支持，即驗證了進階高階級寫作測驗，能有效將華語學習者的寫作表現區分為 CEFR 的 B1 和 B2 兩等級。

本測驗進階高階級測驗總分為書信寫作及觀點論述兩題型的成績加總，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 5 至 7 分者，可取得進階級(Level 3)證書，總分介於 8 至 10 分者，可取得高階級(Level 4)證書。

表 6　進階高階級通過門檻分數

測驗等級	證書等級	分數範圍
進階高階級	高階級	8-10
	進階級	5-7

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)，說明如下：

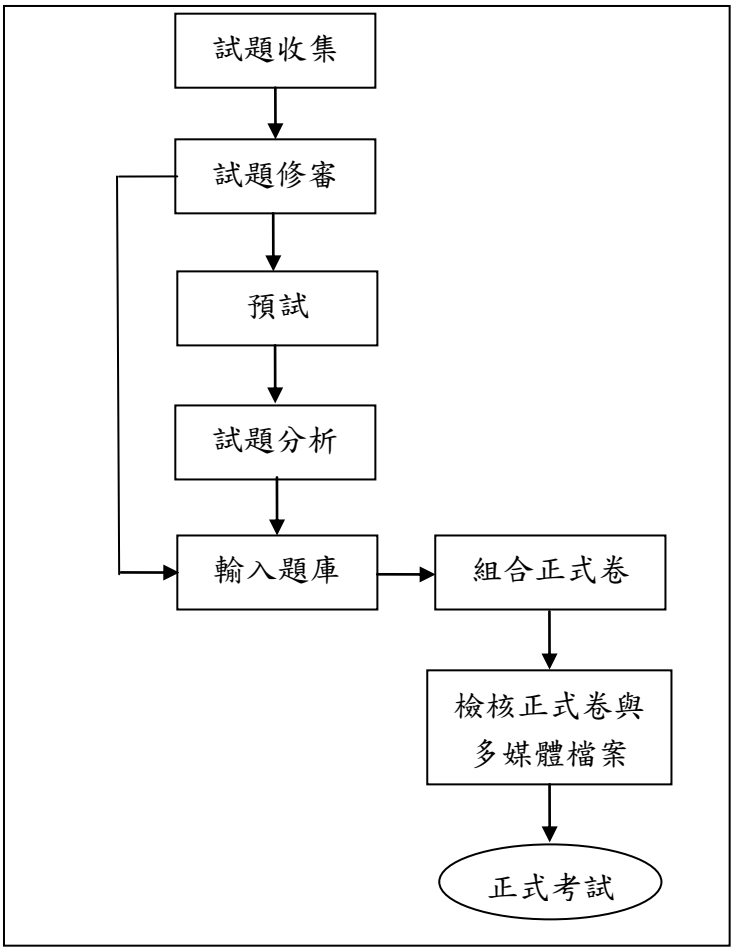

圖 1　正式考試製卷流程

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，本測驗研發人員(以下簡稱研發人員)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) 試題修訂與製作相關多媒體檔案

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，便將題目與考試說明影片匯入考試系統。其中，因應本年度調整測驗架構，本測驗重新製作入門基礎級與進階高階級考試說明影片。

3. 預試

修訂後的試題，一部分輸入題庫，一部分作為預試題目。本會今年度一共舉辦兩場全國性寫作測驗預試，分別是 3 月 31 日進階高階級預試，到考人數為 131 人；7 月 13 日入門基礎級預試，到考人數為 97 人。

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相模式(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相模式分析的 Facets 3.71.3 版的部分給分模式(partial credit model；簡稱 PCM)對資料進行分析，部分給分模式如公式 1 所示：

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

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)之均方差(mean-square)(簡稱 Infit MNSQ)來評估預試試題品質。評估標準為：試題之 Infit MNSQ 數值介於 0.5 至 1.5 者，表示試

題適配，意即試題品質與測驗研發目標一致、試題品質良好。今年度入門基礎級預試與進階高階級預試的試題，在經由試題分析後顯示，其 Infit MNSQ 數值皆落於評估標準內，所有預試試題品質良好。

5. 輸入題庫

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。本年度輸入入門基礎級題庫的題數為 17 題，進階高階級為 18 題。

6. 組合正式卷

舉辦考試之前，研發人員自題庫中選取的兩種題型的題目，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，除研發人員進行檢核之外，亦邀集資訊人員與其他測驗的研發人員進行跨組檢核。其主要目的為共同測試寫作考試系統，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) 本測驗研發人員檢核

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) 跨組檢核

檢核無誤的考題，先由資訊人員製成圖檔，並與說明影片上傳至系統中，再由研發人員登入系統，進行模擬交叉測試，檢核試題的字體大小、間距與版面清晰度，並立即回報資訊人員調整。最後邀集其他測驗的研發人員共同測試考試系統，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「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：

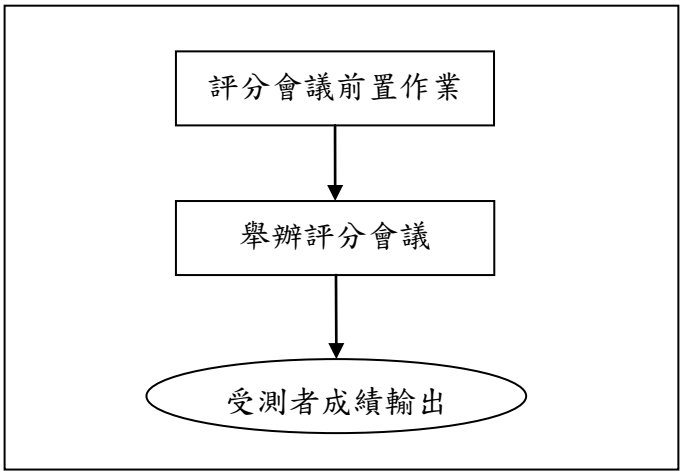

圖 2　評分流程

1. 評分會議前置作業

考試後，研發人員依據該次考試受測者作答反應，針對內容取材的適切性與豐富度，草擬寫作任務評分細則，而後邀請核心教師，即評閱經驗較為豐富的資深評分教師據此進行試評，並提供意見。研發人員再參考所有試評者的建議，修改任務評分細則，並確認結構組織句法表現和詞語表現二向度的評分原則內容，最後依修訂後的評分標準，挑選各級分樣卷，以供評分教師於正式評閱時參考。

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成績，並與核心教師共同討論分數差距較大者，以決定最後的成績。

因應測驗等級架構的調整，本會原先建置的「寫作測驗線上評分系統」進行修改，因此暫時改採電腦文字檔的評閱方式。其作法為等老師們評閱了兩篇之後，便列印評閱結果，若發現評分問題，立即向該位評分老師反應，待釐清相關問題後再繼續評閱。為避免評分標準偏離，於評閱過程中，多次重複上述程序。結束評閱後，由研發人員儲存所有老師的評閱文本，以便會後進行成績彙整與分析。

舉辦評分研習的目的，在於透過實作與討論，以釐清評分概念。本年度本會共舉辦了 12 場評分會議(含研習)。其中，入門基礎級書信寫作題型有三場、進階高階級書信寫作、觀點論述題型分別有四場與五場。

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 2013 年 11 月 3 日所舉辦的進階高階級正式考試之信效度來總結寫作測驗之效能。

本次考試共有 122 名到考，書信寫作題型有四位評分者參與評分工作，評閱篇數為 118 至 120 篇不等，觀點論述題型則由三位評分者進行評分，評閱篇數皆為 120 篇。其中一位評分教師參與二種題型之評分工作，評分教師共有六人。

(一) 信度

所謂信度，指的是測驗結果的穩定性與一致性。一份測驗，若無論在什麼時間、什麼地點，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(即成績)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)。

本節將詳述 2013 進階高階級華語文寫作測驗正式考試之信度。其中，將以「評分者嚴格度變異」來評估評分者內信度，並以「斯皮爾曼等級相關」(Spearman rank order coefficient)進行評分者間信度分析。

1. 評分者內信度

此部分採用 Facets 3.71.3 版的部分給分模式對資料進行分析，檢視評分者嚴格度差異，以及評分者內信度。由於本測驗採分析式評分，以下將分別說明書信寫作與觀點論述題型分向成績之評分者嚴格度結果。

書信寫作題型分向成績的評分者嚴格度如表 7 所示，評分者 B04 給分略為嚴格，嚴格度為 0.260；而評分者 B02 給分略為寬鬆，嚴格度為-0.260。以嚴格度平均值 0 作為標準來看，四位評分者嚴格度均相差±0.3logit 以內，顯示四位評分者評分結果相當一致。在評分教師給分穩定性方面，由評分者嚴格度標準誤所提供的直接證據顯示，各評分者之標準誤介於 0.070 至 0.075 之間。整體而言，四名評分者的標準誤變異情形差異不大，即表示評分者皆具有自身評分的穩定性。而由適配性指標(Infit MNSQ)評估評分者自身給分一致性之間接證據也顯示，四位評分教師均符合評估標準，數值介於 0.5 至 1.5 之間，顯示評分者內一致性佳，給分符合模式預期，評分穩定性良好。

表 7　書信寫作題型分向成績的評分者嚴格度

評分者編號	評閱篇數	觀察的平均值	嚴格度	標準誤 (standard error)	Infit MNSQ
B04	118	3.65	0.260	0.070	0.99
B03	118	3.77	0.052	0.072	1.01
B05	122	3.81	-0.052	0.072	0.93
B02	118	3.93	-0.260	0.075	1.05

註：觀察的平均值表示評分者平均給分成績；Infit MNSQ 表示訊息加權適配度統計量之均方差。

觀點論述題型分向成績的評分者嚴格度如表 8 所示，三位評分者中，評分者 B06 給分較為嚴格，嚴格度為 0.646，其餘二位評分者嚴格度差異不大。以嚴格度平均值 0 作為標準來看，三位評分者中，有二位嚴格度相差±0.3logit 以內，而 B06 嚴格度與另外二位評分者差異較大。在評分教師給分穩定性方面，由評分者嚴格度標準誤所提供的直接證據顯示，各評分者之標準誤介於 0.092 至 0.093 之間，整體而言，三名評分者的標準誤變異情形差異不大，即表示評分者皆具有自身評分的穩定性。而由適配性指標(Infit MNSQ)評估評分者自身給分一致性之間接證據也顯示，三位評分教師均符合評估標準，數值介於 0.5 至 1.5 之間，顯示評分者內一致性佳，給分符合模式預期，評分穩定性良好。

表 8 觀點論述題型分向成績的評分者嚴格度

評分者編號	評閱篇數	觀察的平均值	嚴格度	標準誤 (standard error)	Infit MNSQ
B06	122	2.68	0.646	0.093	0.78
B07	122	3.12	-0.263	0.092	1.03
B05	122	3.18	-0.383	0.093	1.23

註：觀察的平均值表示評分者平均給分成績；Infit MNSQ表示訊息加權適配度統計量之均方差。

2. 評分者間信度

針對書信寫作題型四位評分者評分結果進行斯皮爾曼等級相關分析，以了解兩兩評分者的評分者間信度，結果如表 9 所示。整體級分介於.679 至.751 之間(p<.01)，顯示評分者給分具有中高度正相關；分向成績部分，以「結構組織句法表現」相關係數較高，介於.668 至.825 之間(p<.01)，達中高度至高度正相關，「任務完成度」與「詞語表現」相關係數分別在.520 至.698，以及.552 至.674 之間，具有中度正相關，皆達到.01 之顯著水準。

表 9　書信寫作題型評分者間斯皮爾曼等級相關

組別	評閱篇數	整體級分	任務完成度	結構組織句法表現	詞語表現
B02 vs. B03	118	.716**	.520**	.718**	.552**
B02 vs. B04	118	.679**	.600**	.732**	.599**
B02 vs. B05	118	.719**	.667**	.668**	.674**
B03 vs. B04	118	.745**	.698**	.720**	.658**
B03 vs. B05	118	.751**	.564**	.825**	.649**
B04 vs. B05	118	.733**	.637**	.701**	.668**

**表示p<.01

表10為觀點論述題型三位評分者給分的斯皮爾曼等級相關分析結果，整體級分方面，評分者之間的相關係數介於.606至.631之間，均達到.01的顯著水準，具有中高度正相關；向度分數方面，三位評分者在「形式概念語言能力」給分具有中高度的正相關，相關係數分別為.684、.723與.706(p<.01)，「任務完成度」則較低，相關係數分別為.538、.579及.591，具有中度正相關，皆達到.01顯著水準。

表 10　觀點論述題型評分者間斯皮爾曼等級相關

組別	評閱篇數	整體級分	任務完成度	形式概念語言能力
B05 vs. B06	122	.622**	.538**	.684**
B05 vs. B07	122	.606**	.579**	.723**
B06 vs. B07	122	.631**	.591**	.706**

**表示$p<.01$

由上述評分者信度分析結果可知，2013年寫作測驗正式考試六位評分者皆具有評分者內一致性，自身給分穩定度良好；評分嚴格度方面，書信寫作題四位評分者嚴格度均相差±0.3 logit以內，觀點論述題三位評分者中有兩位嚴格度差異落在±0.3 logit以內；斯皮爾曼等級相關結果，書信寫作與觀點論述題型評分者間的兩兩相關係數達中度或中高度正相關，評分者間信度尚可。

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，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。而在受測者進行測驗時，收集其對自身寫作能力

的主觀評估，進行受測者自評結果與測驗成績之相關分析，則屬同時效度(concurrent validity)，可作為效標效度的一種證據來源。

以下將分別以程序性效度、試題分析及同時效度三方面來描述本次正式考試的內容效度、建構效度以及效標效度。

1. 程序性效度

研發人員針對評分工作制訂的標準化流程，主要分為兩個階段。第一階段為評分會議前置作業，其工作內容為根據試題設定的「寫作任務」，草擬評分細則，邀請資深評分教師進行試評，研發人員再參考試評意見加以修改，並據此挑選各級分樣卷、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。同時，研發人員檢視評分結果之一致性，即時提供回饋。

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 11。

表 11　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。

透過標準化評分流程，寫作測驗評分教師嚴格度雖然有所不同，如書信寫作題型評分者 B04、觀點論述題型評分者 B06 嚴格度偏嚴；然而，二種題型共七人次的評分者中，有六人次評分者嚴格度與平均值相差在±0.5 logit 以內，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. 建構效度

本測驗之組卷方式是依據試題反應理論(IRT)而來。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 版的部分給分模式(如公式 1 所示)對資料進行分析，結果如表 12 所示，書信寫作題型之三個評分向度中，結構組織句法表現較難，任務完成度和詞語表現較為容易；觀點論述題型之任務完成度難度略高於形式概念語言能力，但差異不大。再採用 Infit MNSQ 介於 0.5 到 1.5 的標準做為評估試題是否與單向度試題反應理論模式適配(亦即 Infit MNSQ 超出範圍為題目不符合單向度試題反應理論模式)，結果顯示各向度與模式的適配情形皆良好，二種題型之所有向度皆符合標準，顯示各題型之評分向度測量到相同的潛在特質，也就是寫作表達能力，意即本測驗進階高階級正式考試具有一定程度的建構效度。

表 12　試題向度難度分布

題型	向度	難度	標準誤(S.E.)	Infit MNSQ
書信寫作	結構組織句法表現	0.240	0.060	0.80
	任務完成度	-0.048	0.062	1.20
	詞語表現	-0.192	0.066	0.99
觀點論述	任務完成度	0.157	0.071	1.13
	形式概念語言能力	-0.157	0.082	0.86

註：Infit MNSQ 表示訊息加權適配度均方差。

3. 效標效度

本測驗採用受測者自評寫作能力表現與實際測驗表現之間的關聯性來評估效標效度中的同時效度。於正式考試結束後，請受測者填答一份寫作能力自評問卷(如附件 2)以收集相關資料進行同時效度分析，自評問卷採李克特四點量表(Likert scale)，共有五道試題，受測者在閱讀完每道試題的能力描述後，從「非常困難」、「困難」、「容易」、及「非常容易」四個選項中，圈選出一個最符合的選項。計分方式為：圈選「非常容易」得 4 分；「容易」得 3 分；「困難」得 2 分；「非常困難」得 1 分。五道試題回答結果之加總即為受測者寫作能力自評結果，隨後再分別與其測驗總分、測驗通過等級(不通過標記為 0；通過進階級標記

為 1；通過高階級標記為 2)進行相關分析。

結果顯示，受測者自評結果與測驗總分的積差相關係數為.432(p<.01)，與測驗通過等級的等級相關係數為.400(p<.01)，顯示受測者自評寫作能力與測驗總分和通過等級之間均有正相關存在，自評寫作能力越佳者，其寫作測驗總分越高，通過測驗等級越高。再將各題回答結果與測驗總分、通過等級進行斯皮爾曼等級相關分析，所得結果如表 13 所示。自評問卷中，所有題目的答題反應與測驗總分的相關皆達到.01 的顯著水準，相關係數介於.301 至.479 之間，表示回答越容易的受測者，其測驗總分也越高。而與通過等級的相關，同樣所有題目之相關係數達顯著水準(p<.01)，數值介於.243 至.369 之間，表示回答越容易的受測者，通過測驗等級也越高。

從表 13 可以看到自評問卷中 Q1「『寫一封比較詳細的信，告訴別人自己的經驗』對你來說，難度怎麼樣？」、Q2「『寫一封比較詳細的信，告訴別人自己對一件事情的感覺』對你來說，難度怎麼樣？」與測驗總分及通過等級的相關係數最高，研發人員推測此一結果的原因，可能是由於對一位語言學習者來說，無論是在日常生活中與他人溝通，或是在課堂上所接受的語言訓練中，以上兩種溝通任務較為常見，因此這類型的自評敘述對於受測者來說也較為容易評估，與測驗表現的相關性較高。相較來說，問卷中 Q4「『撰寫闡述論點的文章或報告時，對特定觀點提出支持或反對的理由。』對你來說，難度怎麼樣？」、Q5「『撰寫闡述論點的文章或報告時，適度強調重點和相關細節，有系統地發展論述。』對你來說，難度怎麼樣？」闡述觀點這類型的題目，雖然「對特定觀點提出支持或反對的理由」其難度比「告訴別人自己對一件事情的感覺」高一點，但要闡述自己的看法並非難事，不過針對「若是再適度強調重點和相關細節，有系統地發展論述」，在自評問卷中的測驗總分及通過級分的相關係數就相對低一些，會導致這樣的結果，可能也是受測者在日常生活中或是課堂上較少接觸這一類的練習，因此反應在自評結果上，這類型的問題對於受測者來說較為不易評估，與測驗結果的關聯性略低，但仍達到.01 顯著水準(p<.01)。

表 13　自評問卷各題與測驗總分、通過等級之相關分析結果

	Q1	Q2	Q3	Q4	Q5
測驗總分	.388**	.479**	.336**	.357**	.301**
通過等級	.322**	.369**	.311**	.299**	.243**

註：Q1-Q5 表示自評問卷題號；** 表示 $p < .01$。

2013 年華語文寫作測驗進階高階級正式考試之效度指標顯示，透過標準程序訓練的六位評分教師，在評分上均具有一定穩定度，確保了一定程度的內容效度。所有評分向度皆評量到相同能力，具有建構效度；受測者自評寫作能力與測驗總分、通過等級皆有正相關存在，顯示本測驗具有效標關聯效度。

五、結論

本測驗 2013 年技術報告，首先簡介進階高階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，其次說明標準化的製卷與評分作業流程，最後分析正式考試的信效度，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」來評估評分者內一致性，並以「斯皮爾曼等級相關」進行評分者信度分析；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是經由標準化程序而來，此程序效度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，對於受測者成績來說便相當重要。基於此點考量，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，日後本會在評閱工作結束後，將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，其主要目的在於確認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 2013 年度全國性進階高階級寫作測驗正式考試的信度與效度分析資料來看，可大致總結以下三項要點：

(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。

(三) 受測者自評結果可作為測驗結果的有效預測效標。例如：自評寫作能力較佳者，其測驗表現也較佳。

綜上所述，2013 年度進階高階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。也因測驗架構調整，一等測驗涵蓋兩個等級，更能發揮測驗效能。

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。

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

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

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.

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　進階高階級寫作測驗標準設定研究問卷調查結果

問卷內容	平均數
1. 我了解本次 TOCFL 寫作測驗與 CEFR 對應研究會議的目的。	4.0
2. 會議帶領者對於標準設定方法的流程說明得很清楚。	3.9
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
4. 會議帶領者對於 CEFR B1 與 B2 最低能力描述說明得很清楚。	3.9
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。	3.8
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
7. 在第二回合，提供考生文本級分有助於我判斷 CEFR 等級。	3.8
8. 我是根據 B1 最低能力描述判斷評分原則 B1 的通過門檻分數。	4.0
9. 我是根據 B2 最低能力描述判斷評分原則 B2 的通過門檻分數。	4.0
10.　我對於自己所設定的通過門檻分數(cut score)有信心。	3.6
11.　我是根據 B1 最低能力描述判斷考生文本 CEFR 等級。	4.0
12.　我是根據 B2 最低能力描述判斷考生文本 CEFR 等級。	4.0
13.　我對於自己所判斷的考生文本 CEFR 等級有信心。	3.6

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。

附件 2　進階高階級寫作測驗問卷

1.「寫一封比較詳細的信，告訴別人自己的經驗。」對你來說，難度怎麼樣？

How difficult is it for you to write a letter about your experiences in detail?

○非常困難 Very difficult ○困難 Difficult ○容易 Easy ○非常容易 Very easy

2.「寫一封比較詳細的信，告訴別人自己對一件事情的感覺。」對你來說，難度怎麼樣？

How difficult is it for you to write a detailed letter in which you tell someone else how you feel about something?

○非常困難 Very difficult ○困難 Difficult ○容易 Easy ○非常容易 Very easy

3.「寫一封比較詳細的信，提供朋友關於音樂或電影的訊息及想法。」對你來說，難度怎麼樣？

How difficult is it for you to write a detailed letter about the information and your thoughts regarding some music or a movie?

○非常困難 Very difficult ○困難 Difficult ○容易 Easy ○非常容易 Very easy

4.「撰寫闡述論點的文章或報告時，對特定觀點提出支持或反對的理由。」對你來說，難度怎麼樣？

How difficult is it to “compose an essay or report in which you provide specific reasoning that supports or criticizes a main argument”?

○非常困難 Very difficult ○困難 Difficult ○容易 Easy ○非常容易 Very easy

5.「撰寫闡述論點的文章或報告時，適度強調重點和相關細節，有系統地發展論述。」對你來說，難度怎麼樣？

How difficult is it for you to compose an essay or report in which you appropriately emphasize significant points and relevant supporting details while systematically developing your argument?

○非常困難 Very difficult ○困難 Difficult ○容易 Easy ○非常容易 Very easy

書　　名：華語文能力測驗技術報告—2013(4)
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：2015 年 12 月
定　　價：新台幣 100 元