華語文能力測驗技術報告—2014(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

一、 前言 1
二、 簡介 2
(一) 能力描述 2
(二) 測驗題型 2
(三) 評分方式 4
(四) 評分原則 4
(五) 通過門檻 7
三、 測驗標準化流程 11
(一) 標準化製卷流程 11
(二) 標準化評分流程 15
四、 測驗評估 17
(一) 信度 17
1. 評分者內信度 18
2. 評分者間信度 19
(二) 效度 20
1. 程序性效度 21
2. 建構效度 21
(三) 入門基礎級第二部分題型的評分向度合宜性研究 23
五、結論 27
六、文獻 29

表目錄

表 1　通過等級與能力描述 2
表 2　測驗題型 3
表 3　第一部分評分原則 5
表 4　第二部分書信寫作評分原則 6
表 5　標準設定各回合判斷結果之標準差 9
表 6　入門基礎級通過門檻分數 10
表 7　第一部分評分者嚴格度 18
表 8　第二部分分向成績的評分者嚴格度 19
表 9　第一部分評分者間斯皮爾曼等級相關 19
表 10　第二部分評分者間斯皮爾曼等級相關 20
表 11　標準化評分會議的工作內容 21
表 12　第一部分試題難度分布 22
表 13　第二部分評分向度難度分布 23
表 14 任務完成度、結構組織句法表現、詞語表現對整體級分的逐步迴歸分析摘要表 24
表 15 形式適切度、句內結構正確度、文句銜接適切度對結構組織句法表現級分的逐步迴歸分析摘要表 25
表 16 詞語正確度、詞語簡潔度對詞語表現級分的逐步迴歸分析摘要表 26

圖目錄

圖 1　正式考試製卷流程 12

圖 2　評分流程 15

附件目錄

附件 1　入門基礎級寫作測驗標準設定研究問卷調查結果……...…30

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)，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，為能在區分受測者是否通過某一個等級之餘，更進一步分辨出通過測驗群體其能力的高低，同時提高測驗效能及經濟效益，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。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，受測者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。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。

本報告包含三個部分，首先簡介 2014 年入門基礎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 2014 年入門基礎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。

[1] 預計於 2015 年著手研發流利精通級寫作測驗。

二、 簡介

2014 年度華語文寫作測驗正式考試等級為入門基礎級(Band A)與進階高階級(Band B)，依照測驗成績可細分為入門級(Level 1)、基礎級(Level 2)與進階級(Level 3)、高階級(Level 4)四個等級，分別對應 CEFR 的 A1 (Breakthrough)、A2 (Waystage)、B1 (Threshold)與 B2 (Vantage)。2013 年寫作測驗技術報告中已說明進階高階級寫作測驗相關內容(國家華語測驗推動工作委員會，2015)，因此本年度(2014)技術報告僅針對入門基礎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，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，A1 等級的學習者在寫作表達能力方面，能書寫簡單、不連貫的短語和句子;在整體寫作互動能力方面,能以書面形式索取或是提供詳細的個人資料。

A2 等級的學習者在寫作表達能力方面，則能運用簡單的連接詞，例如：「和」、「但是」、「因為」等，書寫一連串簡單的短語和句子；在整體寫作互動能力方面，能對立即的需求，書寫簡短、格式固定的便條。

綜合比較以上兩個等級的寫作能力描述，可推知 A1 等級學習者尚不具備連結句子完成簡短描述的能力，而 A2 等級學習者則能連結句子，傳達與切身相關常見的任務內容。本測驗依此制定入門基礎級寫作能力描述,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
入門級	能寫出簡單、不連貫的短語和句子。
基礎級	能運用簡單的連接詞寫出簡短的電子郵件，表達立即的需求，如：感謝、道歉、邀請等。

(二)測驗題型

為了在有限的測驗時間內有效測出並區別 A1、A2 兩個等級受測者之寫作能力，本測驗研發人員參考 CEFR 對 A1 及 A2 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與難度適切性，據此制定測驗題型。

從 CEFR 溝通式寫作活動之整體書面表達得知，入門基礎級學習者之寫作能力是由「能書寫簡單、不連貫的短語和句子。」發展至「能運用簡單的連接詞，例如：『和』、『但是』、『因為』等，書寫一連串簡單的短語和句子。」因此，此階段的學習者寫作能力表現可分為「短語或單句層次之寫作能力」與「簡單描述段落層次之寫作能力」兩大特徵。而「短語或單句層次之寫作能力」為入門級學習者之寫作能力表現，「簡單描述段落層次之寫作能力」則較偏向基礎級學習者之寫作能力表現。

據此，本會將入門基礎級寫作測驗題型分為兩大部分，第一部分是「句子重組」、「完成對話」及「圖片描述」，以此三種題型作為評量受測者單句寫作能力之測驗題型。其中「句子重組」之寫作任務為「將題目提供的幾個詞語重新排列，組合成一個正確的句子。」；「完成對話」則為「寫出一個完整的句子，完成對話。」；「圖片描述」要求受測者「根據圖片 A 和圖片 B 的內容，各寫兩個完整的句子。」以檢視受測者是否能根據字詞、對話及圖片的引導寫出完整的句子。

至於第二部分，則選擇以側重評量寫作互動活動的「書信寫作」作為測驗題型，主要由於此一寫作活動最貼近此等級學習者之生活經驗與實際寫作需求。因此，第二大題選擇以「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 70 至 120 個字的私人信件，以檢視受測者是否能以一連串簡單的短語和句子，描述與切身相關的經驗，傳達簡單私人訊息，表達立即的需求。入門基礎級寫作測驗題型如表 2 所示。

表 2　測驗題型

	題型	題數	字數	時間
第一部分	句子重組	2	每題 20 字以內	共 20 分鐘
	完成對話	2		
	圖片描述	4		
第二部分	書信寫作	1	70-120 字	20 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)。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，針對具一定篇幅的段落文本採取分析式評分。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

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所制定的寫作評分規準，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，並諮詢華語文教學與語言測驗相關領域專家學者的意見。入門基礎級第一部分「句子重組」、「完成對話」及「圖片描述」三種題型的級分級距設定為 0 至 3 級分，依照寫作表現的不同，給予 0 至 3 的等級；第二部分「書信寫作」的評分級距設定為 0 至 5 級分。依照入門基礎級兩大部分題型及評量重點的不同，制定出兩套評分原則。

第一部分之「句子重組」、「完成對話」及「圖片描述」三種題型的評量重點分述如下。「句子重組」主要側重考生是否能將題目指定的所有詞語組成一個句子;「完成對話」主要側重考生能否使用基本詞語組成一個簡單的句子，適切地詢問或回應對方;「圖片描述」主要側重考生能否使用基本詞語組成一個簡單的句子，適切地描述圖片內容。此外，句子語序是否正確、有無漏詞情形及增漏字問題，亦為此三種題型之共同評量要點。

第二部分書信寫作的評量重點在於檢視受測者訊息的傳遞是否完整清楚,其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設定。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訊息完整度與內容可讀性，以及是否完成題目

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、文句銜接能力與句內結構的掌握度；詞語表現主要檢視受測者對詞語的掌握程度。第一部分評分原則與第二部分書信寫作評分原則，如表 3、表 4 所示。

表 3　第一部分評分原則

題型 級分	第一大題(句子重組)	第二大題(完成對話)	第三大題(圖片描述)
3	能將題目指定的所有詞語組成一個句子，語序正確，且無增漏字或錯字問題。	能使用基本詞語組成一個簡單的句子，適切地詢問或回應對方，句子完整正確。	能使用基本詞語組成一個簡單的句子，適切地描述圖片內容，句子完整正確。
2	能將題目指定的所有詞語組成一個句子，語序正確，但少部分詞語有誤用、增漏字或錯字問題。	大致能使用基本詞語組成一個簡單的句子，以詢問或回應對方，句子大致完整正確。	能使用基本詞語組成一個簡單的句子，大致適切地描述圖片內容，句子大致正確。
1	部分語序不正確；漏一個或兩個詞語；增漏字或錯字問題嚴重。	語序、詞語、增漏字或錯字問題嚴重。	部分內容與圖片不符；語序、詞語、增漏字或錯字問題嚴重。
0	語序錯誤嚴重；漏三個詞語(以上)；增漏字或錯字問題極嚴重；未作答。	答非所問；未作答。	內容完全與圖片無關；未作答。

表 4　第二部分書信寫作評分原則

級分	任務完成度	結構組織句法表現	詞語表現
5	● 達成所有任務，完整清楚地傳遞相關訊息。	● 形式佳(稱謂、署名俱全，且位置正確；無任意分行；標點適切) ● 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	● 大致達成所有任務，大致完整清楚地傳遞相關訊息。	● 形式大致良好(稱謂、署名俱全，但稱謂未頂格；1 個任意分行；標點大致適切) ● 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	● 尚能回應所有任務，簡單傳遞訊息。 ● 尚能回應所有任務，少數內容無關。 ● 達成大部分任務，少數內容缺漏。 ● 達成大部分任務，少數內容不太清楚。	● 形式尚可(稱謂、署名俱全，但稱謂或署名與內文在同一段落；2 個任意分行；標點尚可) ● 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	● 僅達成少部分任務，傳遞少量訊息。 ● 試圖表達，但內容不太清楚，僅能傳遞少量訊息。	● 形式不佳(缺稱謂或署名，或稱謂署名倒置，或未使用題目設定的名字；任意分行，或文章未完，影響文意完整性；標點不佳) ● 文句銜接不佳。 ● 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	● 僅達成極少任務，僅傳遞極少訊息。 ● 試圖表達，但內容極不清楚，幾乎無法傳遞訊息。	● 形式極差(缺稱謂與署名；僅有開頭；標點極差) ● 文句銜接極差。 ● 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；文體不符 (如：寫成記敘文、全文對話形式等)；全文條列式 (文意不連貫，如清單)；低於 20 字		

(五)通過門檻

本測驗透過標準設定(standard setting)程序，訂出入門級與基礎級之通過門檻。由於給分方式依照題型分為 0 至 3 級分，以及 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，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[2]。

1. 簡介此標準設定之目的與門檻設定的方法。
2. 介紹入門基礎級測驗與 CEFR 架構，並說明依據 CEFR 之 A1 及 A2 等級能力描述所定義之入門級與基礎級最低能力描述(minimum performance level descriptions)。
3. 說明句子重組題型內容與評分原則。
4. 請成員依據提供的入門級、基礎級最低能力描述，分別與句子重組題型之評分原則進行配對，決定入門級和基礎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. 請成員閱讀句子重組題型的 10 篇受測者文本後，依據入門級、基礎級最低能力描述，分別判斷每篇文本的 CEFR 等級(A2、A1、不到 A1)，並寫下判斷依據。
6. 提供成員根據步驟 4 與 5 的判斷結果所得之回饋訊息(Cizek & Bunch，2007)。回饋訊息包含：(1)入門級與基礎級 0 至 3 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 CEFR A2、A1、不到 A1 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。

[2] 因入門基礎級測驗於 2014 年 7 月舉行標準設定會議時，適用於判定 CEFR 等級的標準文本篇數有限，故當次會議先進行各題型評分原則配對部分；另於 2015 年 1 月舉行第二次標準設定會議，完成文本 CEFR 等級判斷部分。

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，設定出入門級與基礎級句子重組題型之通過門檻。

設定入門基礎級完成對話、圖片描述與書信寫作題型之通過門檻時，程序同上述步驟 3 至 10，惟書信寫作題型評分級距為 0 至 5 級分。

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(見附件 1)分為兩個部分，第一部分為針對評分原則配對的調查，共有九題四點量表，平均分數在 3.7 以上；第二部分則為文本等級判斷的調查，共有八題四點量表，平均分數在 3.6 以上。上述結果顯示與會者均同意會議帶領者對會議目的及對標準設定方法的操作流程說明得很清楚、能了解最低能力描述的內容、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(Spearman rank order coefficient)作為依據。標準差部分，從表 5 可知，在入門級通過門檻部分，句子重組、完成對話與書信寫作三種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，圖片描述題則是在第一回合 11 位專家判斷已完全一致，標準差為 0；基礎級通過門檻部分，句子重組與書信寫作題在第一回合標準差較大，第二回合已達到完全一致，標準差為 0，完成對話與圖片描述題則是在第一回合已達完全一致，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；也因此圖片描述題未進行第二和第三回合判斷，而句子重組、完成對話與書信寫作題未進行第三回合。

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
入門級	句子重組	0.522	0.000
	完成對話	0.302	0.000
	圖片描述	0.000	--
	書信寫作	0.467	0.000
基礎級	句子重組	0.522	0.000
	完成對話	0.000	0.000
	圖片描述	0.000	--
	書信寫作	0.505	0.000

由於寫作測驗給分採級分制，由評分教師閱卷後給予等級，屬於等級變項，因此採用斯皮爾曼等級相關分析句子重組、完成對話、圖片描述與書信寫作(共四個題型，各 10 份文本)CEFR 等級判斷，與實際級分之間的關聯性。一般來說，相關係數數值在.7 以上可視為高度相關，.4 以上則視為中度相關。

將不到 A1、A1、A2 分別編碼為 0、1、2，與文本實際級分求相關的結果，句子重組題第一回合的相關係數為.802 至 1.000 (p<.01)、第二與第三回合皆為 1.000 (p<.01)；完成對話題三個回合的相關係數依序為.724 至 1.000 (p<.01)、.836 至 1.000 (p<.01)、.826 至 1.000 (p<.01)；圖片描述以及書信寫作三回合相關係數分別為.772 至.939 (p<.01)、.777 至.939 (p<.01)、.855 至 1.000 (p<.01)，以及.747 至 1.000 (p<.01)、.858 至 1.000 (p<.01)、.858 至 1.000 (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有高度正相關存在，判斷結果與實際得分頗為一致。

華語文寫作測驗入門基礎級標準設定結果，在程序性效度與內部效度二項效度證據均獲得支持，即驗證了入門基礎級寫作測驗，能有效將華語學習者的寫作表現區分為 CEFR 的 A1 和 A2 兩等級。

入門基礎級測驗總分為第一部分句子重組、完成對話、圖片描述題成績加總，分別與第二部分書信寫作成績進行加權後而來，滿分為 49 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 26 至 38 分者，可取得入門級(Level 1)證書，總分介於 39 至 49 分者，可取得基礎級(Level 2)證書。

表 6　入門基礎級通過門檻分數

測驗等級	證書等級	分數範圍
入門基礎級	入門級	26-38
	基礎級	39-49

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：

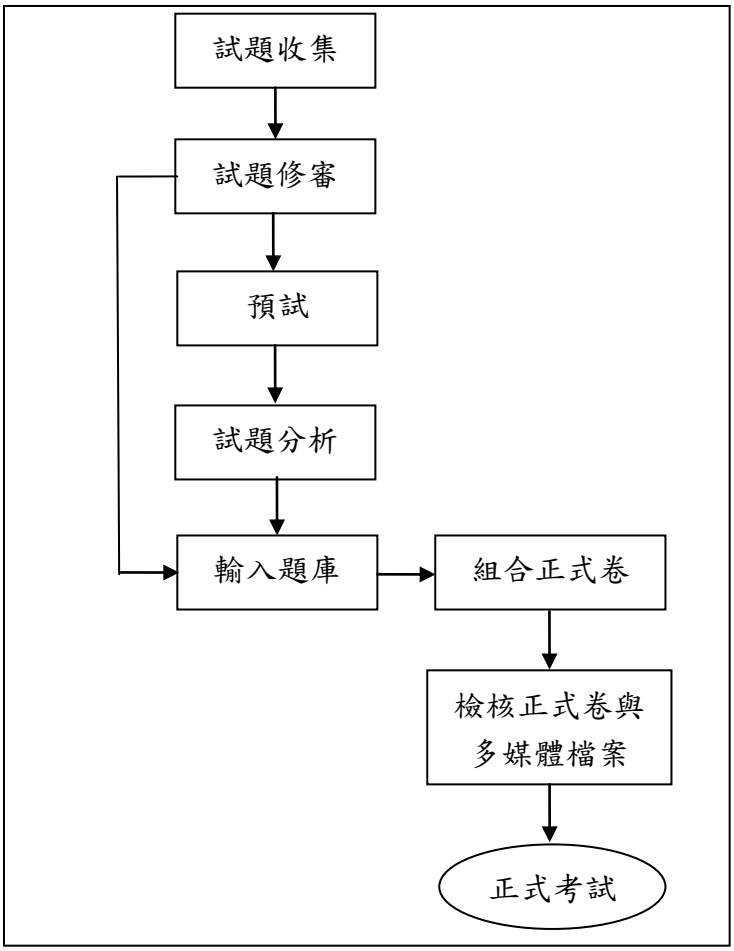

圖 1　正式考試製卷流程

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，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，便將題目輸入題庫，圖片描述題型另請專業繪者依據圖說繪圖及修改。

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。本會今年度舉辦了一場全國性入門基礎級寫作測驗預試，到考人數為 31 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：

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

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(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，意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。本年度(2014)輸入入門基礎級題庫的題數為 39 題；進階高階級為 10 題。

6. 組合正式卷

舉辦考試之前，研發人員自題庫中選取四種題型的題目，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

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考試系統中，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。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：

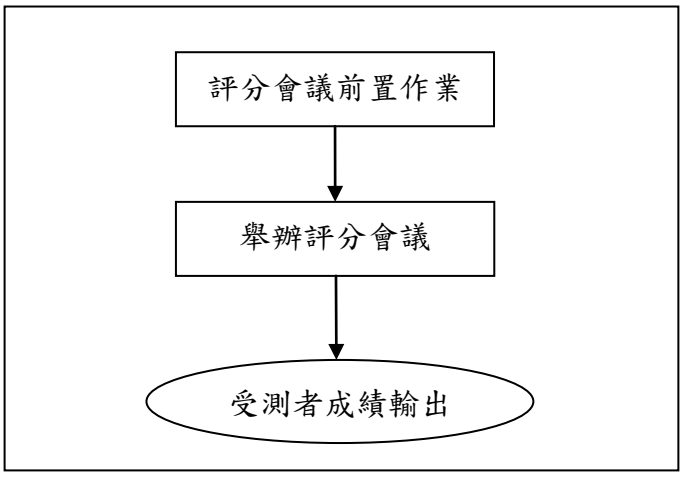

圖 2　評分流程

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。

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 2014 年 11 月 2 日所舉辦的入門基礎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 30 名考生到考，第一部分題型有六位評分者參與評分工作，除了評分者 A13 評閱 239 筆作答反應之外，其餘五位皆評閱 240 筆；第二部分題型則由八位評分者進行評分，除了評分者 A14 評閱 87 篇之外，其餘七位皆評閱 90 篇。而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。

(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 2014 年入門基礎級寫作測驗正式考試之信度，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」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，以及評分者內信度。由於本測驗第二部份採分析式評分，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。

第一部分的評分者嚴格度如表 7 所示，評分者 A13 給分略為嚴格，嚴格度為 0.172；而評分者 A05 給分較為寬鬆，嚴格度為-0.227。以嚴格度平均值 0 作為標準來看，六位評分者嚴格度均相差±0.3 logit 以內，表示六位評分者嚴格度相當一致。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 0.092 至 0.104 之間。整體而言，六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 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，六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　第一部分評分者嚴格度

評分者編號	評閱篇數	觀察的平均值	嚴格度	標準誤 (S.E.)	Infit MNSQ	Outfit MNSQ
A13	239	2.3	0.172	0.092	0.89	0.95
A04	240	2.3	0.138	0.092	0.98	1.04
A02	240	2.3	-0.024	0.097	0.92	0.91
A18	240	2.3	-0.024	0.097	1.04	1.20
A14	240	2.4	-0.034	0.098	1.13	0.85
A05	240	2.4	-0.227	0.104	0.95	0.98

第二部分的評分者嚴格度如表 8 所示，八位評分者中，評分者 A20 給分較為嚴格，嚴格度為 0.665，而評分者 A13 給分略為寬鬆，嚴格度為-0.355。以嚴格度平均值 0 作為標準來看，八位評分者中，有六位嚴格度相差在±0.3 logit 以內，評分標準較為一致；而 A20 評分較為嚴格，A13 評分較為寬鬆。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 0.145 至 0.162 之間，整體而言，八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 Infit MNSQ 與 Outfit MNSQ

評估評分者自身給分一致性之間接證據也顯示，八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　第二部分評分者嚴格度

評分者編號	評閱篇數	觀察的平均值	嚴格度	標準誤 (S.E.)	Infit MNSQ	Outfit MNSQ
A20	90	3.4	0.665	0.145	0.98	1.01
A04	90	3.8	0.019	0.154	0.97	0.91
A02	90	3.8	-0.005	0.155	1.01	1.21
A18	90	3.8	-0.005	0.155	1.01	0.92
A05	90	3.8	-0.029	0.155	1.13	1.12
A19	90	3.9	-0.126	0.157	0.82	0.86
A14	87	3.9	-0.165	0.162	1.21	1.32
A13	90	4.0	-0.355	0.162	0.79	0.75

2. 評分者間信度

針對第一部分六位評分教師給分結果，進行斯皮爾曼等級相關以了解兩兩評分者之間的信度，所有數值介於.585 至 1.000 之間(p<.01)，顯示評分者給分具有中度至高度正相關，皆達到.01 之顯著水準(因篇幅有限，僅呈現所有組合相關係數之平均數)。若比較各題平均的相關係數可發現，數值介於.724 至.973 之間，其中，第 1 題與第 2 題的相關係數最高，分別為.973 與.870，關於此一現象。本會認為上述兩題的相關係數高，是由於該兩題的題型皆為評分較不易受主觀因素影響的「句子重組」。各題評分者間的斯皮爾曼等級相關分析結果之平均，如表 9 所示。

表 9　第一部分評分者間斯皮爾曼等級相關

題號	第 1 題	第 2 題	第 3 題	第 4 題	第 5 題	第 6 題	第 7 題	第 8 題
相關係數平均值	.973	.870	.764	.759	.814	.724	.761	.783
評閱人數	30	30	30	30	30	30	30	30

第二部分書信寫作題型八位評分者兩兩之間的斯皮爾曼等級相關分析結果，如表 10 所示。「結構組織句法表現」的相關係數平均值較高，數值為.796，其次是「詞語表現」，數值為.769，而「任務完成度」的平均相關係數則是.741。

三個向度成績的相關係數平均值，皆達高度正相關。

表 10　第二部分評分者間斯皮爾曼等級相關

向度	任務完成度	結構組織 句法表現	詞語表現	整體級分
相關係數平均值	.741	.796	.769	.805
評閱人數	30	30	30	30

由上述評分者信度分析結果可知，2014年入門基礎級寫作測驗正式考試的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，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相差在±0.3 logit以內；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，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、試題分析結果來描述本次正式考試的內容效度及建構效度。因本次測驗人數較少，故不進行驗證性因素分析。

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，研發人員透過印出之紙本評閱結果監控評分狀況，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 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。

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相差±0.3 logit 以內；在第二部分題目的八位評分教師當中，有六位評分者嚴格度均相差±0.3 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. 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 12、表 13 所示，第一部分題型的 8 道題之中，第 6 題、第 8 題較難，第 4 題、第 3 題較容易，所有題目的難度介於-.536 至.556 之間，再採用 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，結果顯示第 2 題 Infit MNSQ 為 1.53、第 5 題 Outfit MNSQ 為 0.39、第 3 題 Outfit MNSQ 為 1.52，其餘試題數值均符合標準，與模式的適配情形良好。由於 Outfit MNSQ 對於極端的非預期性評分較為敏感；而 Infit MNSQ 則對於累積的非預期性評分結果較為敏感(Eckes, 2009)。因此，許多學者認為 Infit MNSQ 較為適合作為判斷適配度的指標(Pollitt & Hutchinson, 1987；Park, 2004)。故將針對第 2 題進行追蹤，若未來正式考試仍有不適配情形，則考慮停用此題。

第二部分書信寫作題型的三個評分向度之中，任務完成度較難，詞語表現最為容易。再採用 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，結果顯示各向度與模式的適配情形皆良好，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，本測驗入門基礎級正式考試具有一定程度的建構效度。

表 12　第一部分試題難度分布

題號	難度	標準誤(S.E.)	Infit MNSQ	Outfit MNSQ
第 6 題	0.556	0.097	0.88	0.84
第 8 題	0.550	0.094	1.13	1.16
第 7 題	0.391	0.098	0.86	0.90
第 1 題	0.091	0.106	0.98	0.84
第 5 題	-0.188	0.128	0.52	0.39
第 2 題	-0.329	0.114	1.53	1.25
第 3 題	-0.534	0.149	0.94	1.52
第 4 題	-0.536	0.145	0.91	1.02

表 13　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	0.256	0.089	1.04	1.02
結構組織句法表現	0.181	0.098	0.92	0.92
詞語表現	-0.437	0.099	1.00	1.09

(三) 入門基礎級第二部分題型的評分向度合宜性研究

入門基礎級第二部分題型的評分方式採取分析式評分法，先根據考生在「任務完成度」、「結構組織句法表現」、「詞語表現」三個向度的表現給予個別級分，而後透過本會研訂的運算方式給予一個整體級分。其中「任務完成度」主要檢視訊息完整性與內容可讀性；「結構組織句法表現」主要檢視形式、文句銜接、句內結構三個細項的適切度；「詞語表現」則是檢視詞語的正確度與簡潔度。以下將針對三大向度的級分對整體級分的影響與各向度的細項級分對向度級分的影響進行多元迴歸分析，以探討此種評分方式的合宜性。

1. 任務完成度、結構組織句法表現、詞語表現對整體級分的多元迴歸分析

本部分的分析模式採取逐步多元迴歸分析，以任務完成度、結構組織句法表現、詞語表現三大評分向度的級分為預測變項，整體級分為效標變項，其分析結果如表 14 所示。分析結果顯示，此三大向度的級分皆可有意義的解釋整體級分的變異量，三個變項對於整體級分的聯合解釋變異量達到 98.2%。其中，第一個投入的預測變項為結構組織句法表現，此一向度對整體級分的解釋力達 71.6%；第二個投入的預測變項為任務完成度，此向度增加的解釋力為 23.1%；詞語表現為最後投入的預測變項，可增加 3.5%的解釋力。

由逐步多元迴歸分析得出預測整體級分的標準化迴歸方程式如公式 2 所示，從被選入預測變項的 β 值可發現，對於整體級分之預測，預測變項相對重要性由高至低依序為結構組織句法表現(β=.487)、任務完成度(β=.456)、詞語表現(β=.240)。而結構組織句法表現此一評分向度對整體級分影響最高的原因，可能是由於部分考生對該向度之評分細項中的「形式適切度」所涵蓋之中文書信格式與中式標點運用的掌握度較差，導致該細項的得分特別低，進而提高結構組織句法表現對整體級分的解釋力。任務完成度對整體級分的影響力次之，從考生文本可發現，某些考生因未能完整回應所有的寫作任務而影響其成績。至於詞語表現

的標準化係數最低，對整體級分的影響亦最低，其原因可能是受測者在此向度普遍得分較高，由描述統計結果來看，詞語表現的平均數為 4.258，高於其他兩項(結構組織句法表現與任務完成度平均數分別為 3.735 和 3.512)，可印證上述推斷結果。

$$整體級分=.487X_{結構}+.456X_{任務}+.240X_{詞語} \quad (2)$$

表 14 任務完成度、結構組織句法表現、詞語表現對整體級分的逐步迴歸分析摘要表

次序	投入變項	R	R^2	ΔR^2	淨 F 值	標準化迴歸係數	t 值
1	結構組織句法表現	.853	.727	.716	63.895**	.487	14.287**
2	任務完成度	.975	.951	.947	224.875**	.456	13.392**
3	詞語表現	.992	.984	.982	448.982**	.240	6.678**

** $p<0.01$

2. 形式適切度、句內結構正確度、文句銜接適切度對結構組織句法表現級分的多元迴歸分析

結構組織句法表現此一向度的評分涵蓋形式適切度、文句銜接適切度、句內結構正確度三個評分細項。此一部分的分析模式，是以上述三個評分細項的級分為預測變項，以結構組織句法表現的級分為效標變項進行逐步多元迴歸分析，結果如表 15 所示。分析結果顯示，形式適切度、文句銜接適切度、句內結構正確度皆可有意義的解釋結構組織句法表現之級分的變異量，三個變項對於結構組織句法表現之級分的聯合解釋變異量達到 94.0%。其中，第一個投入的預測變項為形式適切度，此一細項對向度級分的解釋力達 83.3%；第二個投入的預測變項為句內結構正確度，此細項增加的解釋力為 9.2%；文句銜接適切度為最後投入的預測變項，可增加 1.5%的解釋力。

由逐步多元迴歸分析得出預測結構組織句法表現之級分的標準化迴歸方程式如公式 3 所示，從被選入預測變項的 β 值可發現，對於結構組織句法表現之級分的預測，預測變項相對重要性由高至低依序為形式適切度(β=.712)、句內結構正確度(β=.276)、文句銜接適切度(β=.176)。形式適切度的標準化係數相對較高，顯示對結構組織句法表現之級分的影響較高，而其原因可能是受測者在此細項普遍得分較低，由描述統計結果來看，形式適切度的平均數為 3.204，低於其他兩細項(文句銜接適切度與句內結構正確度平均數分別為 4.427 和 4.254)，可印證上述推斷結果。

結構組織句法表現此一向度中所涵蓋的三個細項中，對向度影響力最高的是形式適切度，最低的是文句銜接適切度，其原因可能是基礎級的寫作測驗，僅要求受測者能運用基本的銜接詞語，如：「因為……，所以……」、「雖然……，可是……」來回應任務要求即可，因此考生普遍在此項細項能獲得較高的分數，故對結構組織句法表現級分之影響力較低。形式適切度的標準化係數最高，對結構組織句法表現的影響力也是最高，其原因可能是部分考生對形式適切度所涵蓋的中文書信格式或中式標點的掌握度差，因此形式適切度的平均數最低，故對結構組織句法表現之級分的影響力最大。

$$\text{結構組織句法表現之級分}=.712X_{\text{形式}}+.276X_{\text{句內}}+.176X_{\text{文句}} \qquad (3)$$

表 15 形式適切度、句內結構正確度、文句銜接適切度對結構組織句法表現級分的逐步迴歸分析摘要表

次序	投入變項	R	R^2	ΔR^2	淨 F 值	標準化迴歸係數	t 值
1	形式適切度	.916	.840	.833	125.952^{**}	.712	10.757^{**}
2	句內結構正確度	.965	.931	.925	155.719^{**}	.276	5.075^{**}
3	文句銜接適切度	.973	.947	.940	130.580^{**}	.176	2.540^{*}

$^{*}p<0.05$；$^{**}p<0.01$

3. 詞語正確度、詞語簡潔度對詞語表現級分的多元迴歸分析

詞語表現此一向度的評分涵蓋面包含詞語正確度與詞語簡潔度兩個評分細項。此一部分的分析模式，是以上述兩個評分細項的級分為預測變項，以詞語表現的級分為效標變項進行逐步多元迴歸分析，結果如表 16 所示。分析結果顯示，詞語正確度、詞語簡潔度皆可有意義的解釋詞語表現之級分的變異量，兩個變項對於詞語表現之級分的聯合解釋變異量達到 97.6%。其中，第一個投入的預測變項為詞語正確度，此一細項對向度級分的解釋力達 94.7%；第二個投入的預測變項為詞語簡潔度，此細項增加的解釋力為 2.9%。

由逐步多元迴歸分析得出預測詞語表現之級分的標準化迴歸方程式如公式 4 所示，從被選入預測變項的 β 值可發現，對於詞語表現之級分的預測，預測變項相對重要性由高至低依序為詞語正確度(β=.896)、詞語簡潔度(β=.188)。

詞語正確度的標準化係數相對較高，顯示對詞語表現之級分的影響較高，其原因可能是因本測驗之作答方式採鍵盤輸入，部分考生作答時，或許未仔細檢查

並選用正確的詞語，造成錯別字問題嚴重，明顯影響其分數。另外，由於入門基礎級考生需在 70-120 字的字數要求下，回應書信寫作題型的二或三項寫作任務，因此考生在文句上，較少出現冗贅重複的現象，故在詞語簡潔度上，普遍得分較高。由描述統計結果來看，詞語正確度的平均數為 4.096，低於詞語簡潔度(平均數為 4.746)，可印證上述推斷結果。

$$\text{詞語表現之級分}=.896X_{\text{正確度}}+.188X_{\text{簡潔度}} \quad (4)$$

表 16 詞語正確度、詞語簡潔度對詞語表現級分的逐步迴歸分析摘要表

次序	投入變項	R	R^2	ΔR^2	淨 F 值	標準化迴歸係數	t 值
1	詞語正確度	.974	.949	.947	447.163**	.896	26.503**
2	詞語簡潔度	.989	.978	.976	516.753**	.188	5.551**

** $p<0.01$

由以上多元迴歸分析結果得知，本會針對入門基礎級寫作測驗第二部分題型所研訂的三大評分向度與評分細項皆能有效評量受測者的寫作能力，顯示評分原則的向度制定合宜。

五、結論

本測驗 2014 年技術報告，首先簡介新版入門基礎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，最後分析入門基礎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」來評估評分者內一致性，並以「斯皮爾曼等級相關」進行評分者間信度分析；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。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。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。最後並透過多元迴歸分析檢視入門基礎級寫作測驗第二部分題型之三大評分向度與評分細項的合宜性。

由 2014 年度全國性入門基礎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) 本測驗所訂定的評分架構能有效評量受測者的寫作能力。

綜上所述，2014 年度入門基礎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。

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。

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.

Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for Languages: Learning, teaching, assessment (Section H)*. Strasbourg, France: Council of Europe/Language Policy Division. Retrieved from http://www.coe.int/t/dg4/linguistic/Source/CEF-refSupp-SectionH.pdf

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.

Park, T. (2004). An investigation of an ESL placement test of writing using many-facet Rasch measurement. *Teachers College, Columbia University Working Papers in TESOL & Applied Linguistics*, *4*(1), 1-21

Pollitt, A., & Hutchinson, C. (1987). Calibrating graded assessments: Rasch partial credit analysis of performance in writing. *Language Testing*, *4*, 72-92.

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

附件 1　入門基礎級寫作測驗標準設定研究問卷調查結果

	問卷內容	平均數
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。	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。	4.0
	3. 會議帶領者對於本研究 Angoff 法的進行方式說明得很清楚。	4.0
	4. 會議帶領者對於 CEFR A1 與 A2 最低能力描述說明得很清楚。	4.0
	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。	4.0
	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。	4.0
	7. 我是根據 A1 最低能力描述判斷評分原則 A1 的通過門檻分數。	3.9
	8. 我是根據 A2 最低能力描述判斷評分原則 A2 的通過門檻分數。	3.9
	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。	3.7
文本等級判斷	1. 我了解本次 TOCFL 寫作測驗標準設定會議的目的。	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。	4.0
	3. 會議帶領者對於 CEFR 等級與文本配對的進行方式說明得很清楚。	4.0
	4. 會議帶領者對於 CEFR A1 與 A2 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。	4.0
	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。	4.0
	7. 我是根據 A1 及 A2 最低能力描述判斷考生文本 CEFR 等級。	3.9
	8. 大體來說，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。

書　　名：華語文能力測驗技術報告—2014(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：2016 年 11 月
定　　價：新台幣 100 元