華語文能力測驗技術報告—2016（1）

口語測驗信效度

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

附錄

一、 前言

「華語文口語測驗」（以下簡稱本測驗）是一套由「國家華語測驗推動工作委員會」（以下簡稱本會）負責研發，專為母語非華語學習者所設計的口語能力測驗。本測驗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，考量華語學習者的實際口語需求，在命題方面，力求內容之普遍性、真實性，符合一般之交際情境。本測驗施測形式採電腦化測驗，試題透過螢幕和耳機播放，受測者藉由麥克風錄下回答內容並將其回傳至電腦系統。已在 2011 年於臺灣地區推出基礎級與進階級正式考試。

自 2013 年起，本測驗架構調整為三等六級，三等分別為入門基礎級（Band A）、進階高階級（Band B）與流利精通級（Band C），而每一等級又可再依據測驗成績細分為兩級，依序為入門級（Level 1）、基礎級（Level 2）、進階級（Level 3）、高階級（Level 4）、流利級（Level 5）、精通級（Level 6），共六級。此架構相較於僅能區分受測者是否通過測驗而言，能夠更進一步區分出通過測驗的受測群體其能力的高低；同時，對於應試者及試務工作者來說，更符合經濟效益，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，考生即使因為些微分數差距未通過較高等級之門檻，也還有機會通過較低等級之門檻，即一份測驗可同時判斷兩個等級程度。本會已在 2013 年於臺灣地區推出進階高階級正式考試，2014 年推出入門基礎級正式考試，並於 2016 年推出流利精通級正式考試。

本報告分為三部分，首先將針對流利精通級口語測驗的內容、測驗實施與成績公佈之標準化流程進行概述；其次，闡述流利精通級正式考試之信、效度分析結果；最後，根據各項分析結果提出相關討論及建議。

二、 測驗簡介

2016 年度本測驗正式考試等級含入門基礎級、進階高階級和流利精通級，依照測驗成績可區分為入門級、基礎級、進階級、高階級、流利級與精通級，分別對應至歐洲共同語文參考架構（CEFR）之 A1（Breakthrough）、A2（Waystage）、B1（Threshold）、B2（Vantage）、C1（Effective operational proficiency）與 C2（Mastery）。入門基礎級與進階高階級口語測驗之能力描述、測驗題型及題數、評分規準（rubric）與通過門檻等方面的研發成果已分別於 2013 及 2014 年口語測驗技術報告中詳盡說明（國家華語測驗推動工作委員會，2015、2016）；以下針對流利精通級口語測驗之相關內容進行說明。

（一） 能力描述

CEFR 就各等級語言學習者和使用者的口語能力表現，制訂出一套系統性的口語能力描述總表，包含口語的表達能力、互動能力及溝通策略等不同面向。

其中，在表達能力方面，C1 等級學習者能針對自身領域外抽象、複雜、稍具爭議性的主題，發表清楚流暢且組織完整的言談，其組織架構，連接詞語及連接手法，均掌控得宜，也能熟練並自由運用大量詞彙，辨認並運用許多不同的慣用語與口語用法，並隨時以委婉用語，透過迂迴曲折的語言形式來表達思想、交流訊息，以達到得體合宜的交際溝通目的，如觸及避諱、禁忌或禮貌問題的情況；C2 等級學習者則能針對爭議性較高且較具針對性的觀念性議題，即無法單純以對錯二分，需多方考量政經、社會文化背景等種種具體或抽象因素的各類議題，十分精確、適切得體且輕易地運用多樣的修飾詞語手法，嫻熟並自由地運用各種慣用語及流行口語，表達更精細的語義層次甚至是言外之意，產出深度、廣度兼具的論述。

而在互動能力與溝通策略的部分，C1 等級學習者能幾乎不費力地流暢表達自己的思想，只在爭議性高、抽象且複雜的觀點主題上，才會發生因思索而短暫停頓、中斷或修整詞語等言語不平順的情況，也能選用適宜的詞語，開啟話題，發展個人論點，且能吸引他人關注自身評論內容並進而接續話題進行相關內容的討論或爭論；此外，尚能在多輪對話中，維持流暢的互動言談，且一貫維持個人立場及論點；C2 等級學習者則能游刃有餘地進行多元觀點的論述或整合摘要，

還能針對互動過程中的歧異性觀點或複雜的抽象概念，不被察覺地重新構思並進一步做出連貫性的重新論述及說明，加強個人論點的說服力。

本測驗研發人員（以下簡稱研發人員）綜合上述 CEFR 針對 C1、C2 等級所提出的口語能力描述，訂出流利精通級口語測驗各等級通過者所應具備的口語能力，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	在正式場合中，針對複雜、抽象、不熟悉的話題： 1. 能清晰、仔細地描述細節，整合次要主題並做出適當的結論。 2. 能流利、自如、適當且具說服力地回應相反論證的內容。 3. 能使用其他說明、理由及相關例子，延伸並支持自己的論點。
精通級	在正式場合中，針對複雜、抽象、不熟悉的話題： 1. 能發表清楚、流暢、結構完整且邏輯清晰的談話，並能幫助聽者掌握重要的部分。 2. 能遊刃有餘地以清楚、具說服力的論證維持立場。 3. 能彈性地調整說話方式以符合聽者需求。

（二） 測驗方式

根據前述口語能力說明，考量流利精通級口語測驗為華語文口語測驗三等六級測驗中的最高等級，此等級考生應具有整合各種資料來源及充分表達自我意見的能力；因此評量考生是否能針對抽象、複雜、爭議性高的嚴肅主題，進行長時間的意見發表、觀點回應與資料整合，是流利精通級試題的設計方向。為實踐試題設計方向，除期盼能符合國際上各大語言測驗發展的趨勢之外，也能符合實際生活互動溝通的真實情境及需求，是以在測驗方式的研發過程中，審慎評估了 OPI（Oral Proficiency Interview）、SOPI（Simulated Oral Proficiency Interview）、COPI（Computerized Oral Proficiency Interview）等不同測驗方式運用於流利精通級口語測驗的可能性及其優劣：

OPI 屬於直接式（direct）測驗，其測驗形式為考官與考生面對面施測，其優點為考試情境較 SOPI、COPI 自然，考官與考生互動過程中，不僅考官可以獲取考生語言以及非語言的訊息（non-verbal communication），也可根據考生程度、經驗，即時調整考題內容及難度，減低考生因挫折而影響實際口語輸出能力的展

現，考生更有機會在溝通發生問題時使用補救策略，例如澄清所說的話或換句話說以傳達訊息（negotiation of meaning），促使溝通任務的完成，滿足語言最重要的典型功能，也就是兩人或兩人以上的互動溝通（Thrasher，2000）；其缺點則如考生互動表現可能受考官影響，且考官品質良莠不一，或表現在提問技巧，或表現在評分嚴格度及一致性等因素，進而影響測驗信度，是以若採取 OPI 模式，考官訓練為測驗過程的重要環結（Messik，1996），雖可藉由舉辦長期且嚴謹的考官訓練，透過逐次篩選，提升專業知能，但期間所需付出的時間、場地成本較高，且不適合進行大規模且高頻率的口語測驗；此外，OPI 模式的囿限尚包含單一考官無法一次施測多位考生，需訓練較多評分員才足以處理 OPI 考試模式及可能日漸增多的考生（Bachman & Savignon，1986）。

SOPI 與 COPI 皆屬於間接式（indirect）測驗，即以錄音帶或電腦預錄試題的方式測驗考生口語能力，其優勢為具備效率與實用性，可同時對大量受測者施測，也可克服專業考官不足的限制，另有利於推動海外施測；此外，更因所有受測者接受到的試題內容、呈現品質與回應對象皆為一致，且評分方式為試後由專業評分者針對考生錄音內容進行評分，可更進一步客觀落實評分的一致性和公平性，達到測驗之信度與效度。COPI 與 SOPI 雖同屬於間接式測驗，但其差別除了測驗工具為錄音帶或電腦之別以外，尚包含 COPI 的測驗方式為電腦會自動將考題難易不同的題目串在一起，以測出考生實際語言能力，而不需分等施測，且可達到試題多樣化的革新；不過，COPI 目前尚未被運用於大規模口語測驗，主因有二，其一為無法實踐口語互動溝通的功能，其二為考生可自行決定準備時間（張濤，2009）；但也有研究認為，COPI 是一種可行且有效的測試形式，能較真實反映出考生的口語能力、減輕面試過程中的焦慮作用，且電腦化測試不僅在測試形式和內容設計上更具靈活性、豐富性，在評分的準確性和試務行政實施的效率上都優於其他口試形式，能夠滿足一次對大量考生施測的需求（Nuessel，1991；Stansfield，1989；高丙梁，2007；張蓉，2008、2010）。

綜上所述，上述測驗方式各有其優缺點，審慎評估後，研發人員最後選擇比照入門基礎級、進階高階級口語測驗，仍採用以電腦為媒介的 SOPI 做為流利精通級口語測驗的測驗形式，而非 OPI 模式；原因如下：流利精通級口語測驗做為最高等級之口語測驗，互動溝通、摘要能力皆為此階段具代表性的口語能力，

雖前人研究認為 COPI 無法達到口語互動功能（張濤，2009），但若採用以電腦為媒介的 SOPI 模式，善用電腦化測驗於試題設計上具備的靈活性、豐富性，透過限定主題並引導答題方向的題組設計內容，仍可模擬真實互動情境，讓考生藉由回應能力的展現，實踐口語互動溝通功能；摘要能力亦同，電腦化測驗也可透過如以 FLASH 外掛功能呈現文章內容，並於系統中限定閱讀時間、回答時間的方式，模擬真實情境，考察受測者的口頭摘要能力。此外，與 OPI 模式相比，採取間接式測驗，不僅回歸「華語文口語測驗」屬於「標準化語言能力測驗（standardized proficiency test）」之本質，也兼顧信度與效度，更具備效率與實用性。再從信度與效度的面向來看，前人研究指出直接式與半直接式口語測驗的結果相差無幾，如 Clark 和 Li（1986）對 32 名學生進行中文 SOPI 考試，與 OPI 相比，相關係數達.93；又如 Kenyon 和 Tschirner（2000）研究表明 OPI 和半直接式考試的最終給分一致性達到 90%；然而，更有其他相關研究表明，與 OPI 相比，因半直接式測驗可有效減少考生跟考官面對面回答時的應考焦慮因素，考生對 SOPI、COPI 的態度反應更好一些，也更能展現真實口語能力（蔡基剛，2005；高丙梁，2007）。

（三） 測驗題型及題數

依據表 1 流利精通級口語測驗能力描述的內容，大致可將此階段語言學習者的口語表達能力分為「摘要能力」、「回應能力」以及「論述能力」三大面向。其中，流利級與精通級的第一條能力描述皆指向「摘要能力」；流利級與精通級的第二、三條能力描述則共同涵括「回應能力」與「論述能力」。

據此，在流利精通級題型架構的設計上，針對「回應能力」、「摘要能力」以及「論述能力」分別規劃了回應類、摘要類及論述類等三大題型；並考量該階段語言使用者已能擷取各類口頭或書信等不同來源的特性，依據試題內容的素材，區分為透過聽力素材呈現的回應類試題，以及透過閱讀素材呈現的摘要暨觀點論述類試題。前者目的除用以考察考生之回應能力，也盼能達到類似 OPI 測驗模式的互動溝通功能，因而採取類似多輪對話的辯論模式，設計為一組由角色扮演題（Role playing）一題及觀點回應題（Responding to viewpoints）兩題所組成的題組，先以角色扮演題做為議題開展之始，受測者於此題中會先看到一段真人模

擬的新聞播報影像，其後需依新聞內容提供的線索，由試題指定之角色的立場開啟言論，發表與該議題有關的個人觀點，隨後再透過兩題立場與受測者相反，觀點涉及次要主題，且深度、廣度層層遞進的觀點回應題，展現其回應能力；至於摘要暨觀點論述類試題，則因應語言使用者已可對書面信息進行摘要的特點，設計為一組以主題導向，引導語言使用者對不熟悉、複雜、抽象議題產出摘要能力、論述能力的題組，分別以文章摘要題（Summarizing articles）及觀點論述題（Expounding viewpoints）方式呈現；受測者於此題中將先看到總字數約 1000 字左右的文章，於整合文章要點進行口頭摘要後，再針對兩題主題複雜、具爭議性，且延伸自文章要點，但答題可能性不受文章理解程度影響的觀點論述題，進行個人立場與觀點的闡述。

此外，同時考量學習者的需求、動機、特性與可用的語言資源，訂出不同領域中可完成的口語任務，著重於提供在正式場合中可能面臨的各類複雜、抽象、不熟悉議題，受測者在此情境下，需透過整合多方立場、觀點後，以系統性、組織性的論述內容，流利且精準地有效傳達個人立場和觀點。

角色扮演及觀點回應類的題型著重於受測者能否「依據題目設定的情境和角色，使用適當的說話方式，回應相反論證，並在發表意見、看法以及做出結論的過程中，維持自己的立場」。文章摘要類題型則著重於評量受測者能否「依據題目提供的文章，歸納主題及其相關子題的多方重點，並整合出一段持續性且融合多方觀點的統整性口頭摘要」，觀點論述類題型則為延伸自文章摘要題內容的論述類試題，著重於「考察受測者是否能依據題目設定的複雜、抽象及不熟悉主題發表多元化的觀點，提出深度、廣度兼具的論證並維持立場」。

而在受測者正式答題初始，為了讓受測者熟悉測驗方式，另設計了一題試題難度、主題領域複雜性、抽象性及爭議性皆介於高階級和流利級之間的論述能力類試題—「陳述意見題」，該題亦為計分題。流利精通級之題型與題數分布分別如表 2 所示。

表 2　測驗題型與題數分布

測驗等級	題型	題數
流利精通級	陳述意見	1
	角色扮演	1
	觀點回應	2
	文章摘要	1
	觀點論述	2

另外，在作答時間的制定上，本測驗參考了 Main Suite 劍橋主流英語認證、托福（TOEFL）、雅思（IELTS）、全民英檢（GEPT）、漢語水平考試（HSK）、AP 中文考試、法國 DELF 法語鑑定文憑、美國外語教學協會（ACTFL）及美國外交學院（FSI）個別研發的口語測驗（Oral Proficiency Interview，簡稱 OPI）等語言能力測驗中關於準備時間與回答時間的規定，並透過本會舉辦的流利精通級口語能力研究計劃、全國性口語能力測驗預試，分析所有受測者在各種測驗題型的回答時間及內容完整度，最後制定出流利精通級測驗各題型中，除陳述意見題因主題領域偏向一般性、普遍性且較不具爭議性的範疇，試題難度較低且試題內容篇幅較短，準備時間為 1 分鐘，回答時間為 2 分鐘之外，其他題型的每題作答時間皆為 3 分鐘，準備時間則因各類試題素材之別，角色扮演題涉及受測者聽力理解能力而為一題 5 分鐘，文章摘要題涉及閱讀理解能力而為一題 10 分鐘，觀點回應及觀點論述題則為每題 2 分鐘。

（四） 評分規準

口語測驗因受測者的回答內容為開放性的語言輸出，為避免過於主觀性的評分過程影響了受測者能力判定的結果，因而需制定一套可靠實用的評分規準。制定評分規準（或稱原則）時，研發人員考量了各等級測驗評量的重點、語言能力表現的特性、語言任務性質的差異等因素，將評分規準的評分重點分為「內容組織」、「表達能力」、「語言運用」等三個向度。

「內容組織」考察的是任務完成度、話語的組織性和連貫性；其中，任務完成度與口語任務的類別有關，在流利精通級中，語言使用者口語能力的發展重心為說明性能力，隨著可處理情境的多樣性而可細分為摘要類型、觀點回應類型以及觀點論述類型等三類語言任務，各自反映的就是受測者的摘要能力、回應能力

與論述能力。

其中，摘要類題型的「內容組織」向度著眼於「能否整合多方觀點並做出適當結論」，回應類題型的「內容組織」向度則著眼於「能否從不同角度提出具說服力的論證以回應相反論點並維持立場」，論述類題型則著眼於「論點是否清楚明確，且內容詳細」。承上所述，回應類和論述類能力其實都被涵括在論證說明能力的大範圍之內，摘要類能力著重的要點也與論證說明能力發展至巔峰時的「能針對各類次要主題進行整合並做出清晰、仔細的細節描述，並以適當結論結尾」高度相關；也就是說，這三類題型各自指涉的「摘要能力」、「回應能力」以及「論述能力」皆屬於「說明性能力」下的細項能力，這三類題型在內容組織向度中的測驗目標相同，只是採用了不同素材、不同任務類型，希冀能從深度、廣度兼具的角度平衡考察受測者的能力，因此這三大類題型實可共用「內容組織」向度的評分要點，而無需各自規劃對應的內容。

「表達能力」考察的是受測者的語音表現、詞語在句內或句間的重複次數、停頓時間以及語速；「語言運用」考察的則是詞彙語法的適當性、準確性及多樣性。上述各項語言特質為語言學習進程的共性，不受口語任務類別影響，因此可共用於摘要類、回應類和論述類這三大類題型。

據此，研發人員針對流利精通級各類測驗題型規劃了一套可共用之評分規準，並邀請華語教學、能力指標、語言測驗等相關領域的專家學者，根據各等級的口語能力指標、任務型口語的理念、不同主題情境的特性與受測者在口語能力表現的偏誤（如，詞彙、語法）等方面，共同制定出流利精通級口語測驗評分原則，如表 3 所示，適用於流利精通級各類題型，含「陳述意見」、「角色扮演」、「觀點論述」、「觀點回應」及「文章摘要」等題型。

本測驗採整體式評分（Holistic Scoring），評分級距皆設定為 0 至 5 級分，評分者聆聽受測者的音檔內容後，再依據評分原則，給予一整體分數。

表 3　流利精通級評分原則

級分	內容組織	表達能力	語言運用
5	論點清楚，內容豐富詳細，條理清晰。 組織良好，前後連貫，並能做出結構完整且適當的結論。 能依情境適當地說話，整合多方觀點。 從不同角度提出具說服力的論證，並維持鮮明的立場。	表達順暢、流利，幾乎沒有停頓；語音清楚、正確，都能被聽者理解。	具備多樣的詞彙、語法結構，能適當且準確地使用，幾乎沒有錯誤。
4	論點清楚，內容大致豐富詳細。 話語有組織，前後大致連貫，並能做出大致適當的結論。 大致能依情境適當地說話，整合多方觀點。 且大致能從不同角度提出具說服力的論證，並維持立場。 少有重複說明的情況。	表達順暢、流利，少有停頓；語音清楚、正確，都能被聽者理解。	具備多樣的詞彙、語法結構，能較適當、準確的使用，僅有少許錯誤。
3	論點清楚，內容詳細。 話語有組織，前後大致連貫，尚能做出適當的結論。 尚能依情境適當地說話，大致能整合多方觀點。 尚能從不同角度提出具說服力的論證，並維持立場。 偶有重複說明的情況。	表達順暢、尚稱流利，仍有一些停頓；語音清楚、正確，都能被聽者理解。	具備多樣的詞彙、語法結構，幾乎都能適當使用，仍有一些限制、錯誤。
2	論點大致清楚，內容尚稱充足。 話語大致有組織，前後尚稱連貫。 提出的理由，大致能支持自己的論點。 常有重複說明的情況。	語速適中，偶有停頓；語音大致清楚、正確，大致能被聽者理解。	具備足夠的詞彙和語法結構，能大致適當使用，偶有不影響溝通的錯誤。
1	論點不甚清楚，內容不足。 組織較差。 提出的理由不足以支持自己的論點。	語速稍慢，常有停頓；語音大致清楚，不時有錯誤，尚能被聽者理解。	具備足夠的詞彙和語法結構，尚能適當使用，時有錯誤，有時無法直接表達較複雜的意思。
0	考生靜默，沒回答；離題；未依題目要求回答。		

（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。由於流利精通級口語測驗給分方式為 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（Impara & Plake，1997），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. 說明流利精通級各類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，與流利精通級評分原則進行配對，決定流利級和精通級口語最低能力表現最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 判斷結果所得之回饋訊息（Cizek & Bunch，2007），即流利級和精通級 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 所示。
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 所示。
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之通過門檻。

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，標準設定會議過程中均按照既定步驟進行，且在各回合判定後皆給予與會者充分的時間進行討論與分享，如第一回合判定後的討論約進行了一至兩個小時。會議後的問卷調查結果（見附錄一）為所有題目的平均數都達到 3.83 以上，同意百分比皆高達 95%以上。此一結果顯示與會者多數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，以上皆顯示標準設定結果具有程序性效度。

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 4 可知，流利級通過門檻部分，標準差在第一回合最大，經討論後，12 位專家判斷結果於第二回合即已達到完全一致，標準差為 0。精通級通過門檻的標準差也呈現相似情形，於第一回合最大，經討論後，專家判斷趨於一致，於第二、三回合皆達 0.302。

表 4　標準設定各回合判斷結果之標準差

通過等級	第一回合	第二回合	第三回合
流利級	0.539	0.000	0.000
精通級	0.505	0.302	0.302

流利精通級口語測驗標準設定結果，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口語測驗能有效將華語學習者的口語表現區分為 CEFR 的 C1 和 C2 兩等級。

流利精通級口語測驗的計分題目共有七題，各題均採 0 至 5 級分的評分級距，考生測驗總分為七題計分題的成績加總，滿分為 35 分。根據標準設定研究結果，各等級通過分數範圍如表 5 所示。測驗總分介於 21 至 34 分者，可取得流利級（Level 5）證書，總分為 35 分者，可取得精通級（Level 6）證書。

表 5　流利精通級口語測驗通過門檻分數

測驗等級	證書等級	分數範圍
流利精通級	精通級	35
	流利級	21-34

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，就必須制訂一套標準化（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）。口語測驗屬於「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），必須針對其題庫建置與評閱方式，周延規劃具公信力的「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，才能確保口語測驗具有理想的信度與效度。

2016 年度本測驗標準化流程共包含兩部分。第一部分為正式考試製卷流程；第二部分為評分流程。茲分述如下：

（一） 正式考試製卷流程

正式考試製卷流程共包含七個步驟：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，如圖 1 所示。各步驟如下所述：

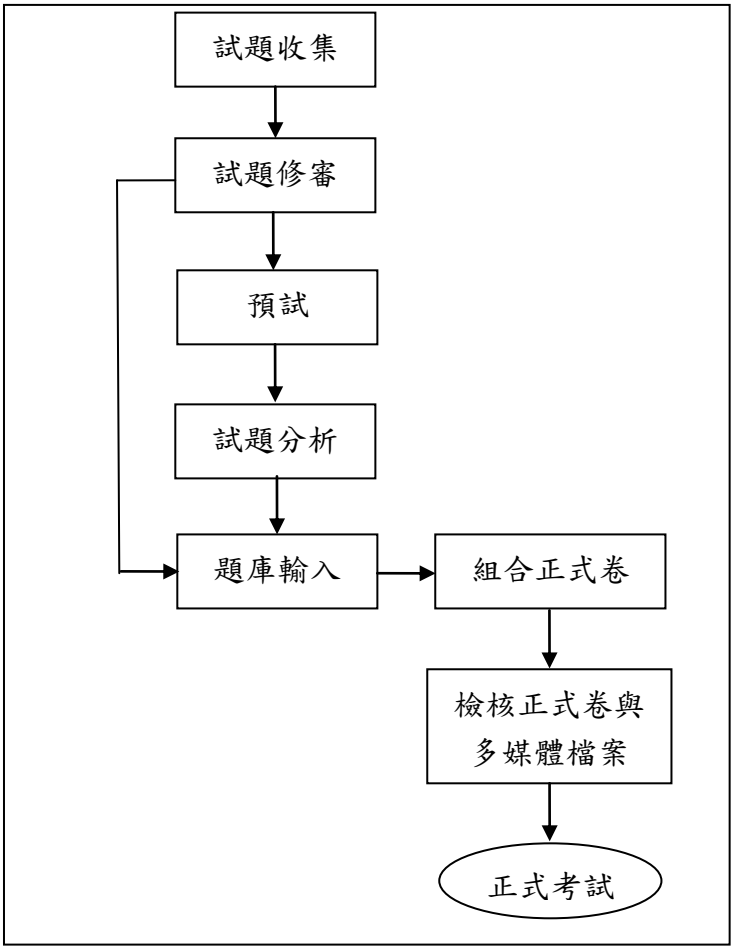

圖 1　正式考試製卷流程

1.　試題收集

2016 年度本測驗流利精通級命題者共計三位，每位命題者每期繳交一套試題，每期命題時長約為六個月。命題者在正式命題前，均已參加本會舉辦之口語測驗命題研習，以充分了解口語測驗之命題方向、口語能力描述與測驗題型等相關內容。同時，研發人員亦提供命題者命題指導文件及《華語八千詞詞表》[1]。根據年度統計，2016 年度流利精通級的命題回收數量為 41 題。

2.　試題修審

本測驗由於各等試卷需涵蓋兩個等級，審查時特別著重於試題難度的適切性，以期兼顧較低等級考生能瞭解並回答問題，而較高等級考生仍能發揮其口語能力。

[1]《華語八千詞詞表》的資料詳見華測會官網：
http://www.sc-top.org.tw/download/8000zhuyin.zip

(1) 會內初審

待命題者繳交試題後，即由研發人員進行第一階段初審工作，隨後回覆審題意見，命題者再根據審題意見修改試題，流利精通級的修改時間約為一個月。

(2) 會內複審

將第一階段會內初審修改後之試題，送交會內非口語測驗之研發人員（約三至五位）進行第二階段試題複審工作，並提供審題意見。其後，由研發人員根據複審意見修改試題，修改時間約為二週。

(3) 專家學者外審

邀請四至五位華語文教學及語言測驗相關領域之專家學者，針對第二階段會內複審修改後的試題，進行第三階段試題審查，並提供審查意見，審查重點著重於試題難度的適切性、試題主題領域的多元性及廣泛性以及試題描述能否有效引導考生產出口語表達內容。外審時長約為三週。最後，再由研發人員依據專家學者的建議修改試題，修改時間約為二週。

(4) 製作試題相關媒體檔案

製作定稿試題之相關媒體檔案，包含拍攝試題影片與後製，及製作圖片、動畫影片和說明影片等。

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為，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口語能力。本會於 2015 年 7 月舉辦全國性流利精通級口語測驗預試。到考人數為 52 名。

4. 試題分析

經過預試階段之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。由於本測驗受測者成績乃經由評分者人工判定（詳見 P.16 評分流程），因此，受測者成績除了受到受測者具備的口語能力及試題難度的影響之外，還受到評分者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之多面向 Rasch 測量模式（many-facet Rasch measurement）（Linacre，1989），對預試資料進行分析。由於計分採級分制（流利精通級為 0-5 級分），屬多元計分方式的試題，因此，本測驗使用可分析多面向 Rasch 測量模式之 Facets 3.71.3 版的部分給分模式（partial

credit model，簡稱 PCM）（Linacre，2013）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（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，口語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 及 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，以口語測驗為例，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。

2015 年 7 月所舉辦之流利精通級全國性預試共計七道試題，試題分析結果顯示，所有試題之 Infit MNSQ 及 Outfit MNSQ 數值皆落於評估標準內，表示所有預試試題品質均為良好。

5. 題庫輸入

本測驗採用開放式題型設計，測驗試題沒有標準答案。評分時，主要依據受測者所回答之內容是否符合測驗研發目標，即在特定語境下，藉由口說，能有效地傳遞訊息、完成溝通任務。故口語測驗題庫之試題來源可分為兩種：經步驟 2 修審程序完成之試題，此其一；經步驟 3 預試後，試題適配度介於 0.5 至 1.5 之間，且評分較無歧異的試題，此其二。2016 年度輸入題庫之流利精通級試題數量為 31 題。

6. 組合正式卷

本測驗正式考試用卷係由進入題庫之試題所組成，組卷時依題型架構及題數

自題庫中選取所需試題；選題時，需考慮試題難易度平均分配於組卷內容中，且試題呈現順序以由易至難為原則；此外，同一份試卷內容不可集中於某一主題，需涵蓋不同主題，以平衡測驗內容，且試題所設定的情境與任務需避免和近幾年的試題重複。

7. 檢核正式卷與多媒體檔案

研發人員需於每次施測前二至四個月將正式卷中所有試題影片上傳至口語考試系統，並登入系統進行模擬交叉測試，模擬測試之檢核重點包含：說明影片內容及語言版本是否正確、試題播放順序是否無誤、考試完畢後音檔存放是否完整、考試進度調整功能是否正常、受測者資料修改等相關功能是否穩定。待上述檢核項目確認無誤後，即完成考試系統測試。

（二） 評分流程

本測驗評分流程主要包含三個步驟，如圖 2 所示。各步驟分述如下：

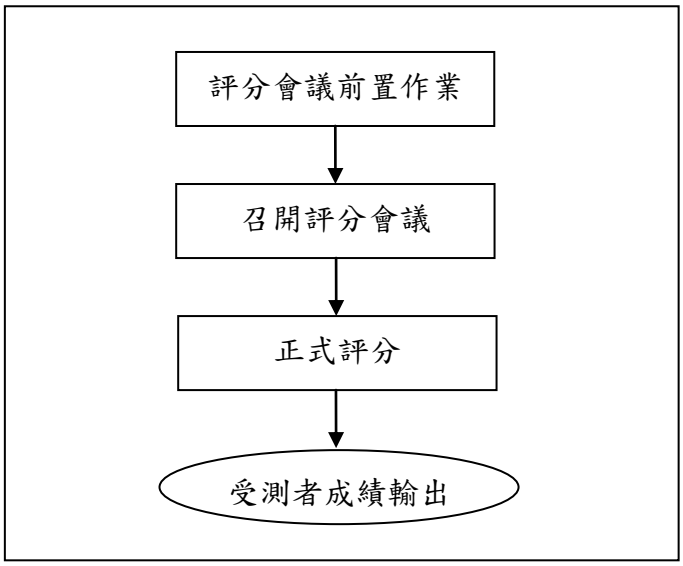

圖 2 評分流程

1. 評分會議前置作業

召開評分會議前，由研發人員挑選該次考試受測者答題音檔，供評分會議試評討論時使用。各題型分別挑選一至二題，每題挑選 15 至 20 個音檔，其中尚需選出各級分之標準音檔，做為評分者熟悉各級分標準之參考依據。

2. 召開評分會議

評分會議召開的主要目的在於調校評分者的評分標準。評分者透過音檔試評與討論，可即時調整評分標準，確保能切實掌握評分要領，以期達到評分一致性。2016 年度，因應 11 月舉辦之流利精通級正式測驗所召開的第一次口語測驗評分會議，共有二位評分者參加。

3. 正式評分

評分會議後，評分者即透過線上評分系統進行為期兩週的正式評分工作，針對受測者每一道試題之回答音檔，獨立進行評分工作，每位受測者的音檔分配給兩位評分者進行評分，評分方式採整體式評分，對考生口語表現的品質，參照評分規準中對各級分表現的整體綜合描述，給予一個整體性的等級分數，評分時間約為兩週。

4. 受測者成績輸出

在評分者完成正式評分並繳交當次評分結果及其評分依據說明後，本會即彙整各評分者之評分結果與依據，針對每一位受測者進行最終成績評定工作；當同一位受測者，兩名評分者評定級分不一致時，再交由第三位評分者評定成績。成績處理上，則以所有評分者給分之眾數做為最終成績；若未能取出眾數，則交由研發人員加入判定最後成績。採用此給分方式，較能避免「針對同一名受測者，不同評分者評分結果差異過大」的現象。

完成受測者成績評定後，研發人員彙整該次正式評分中評分者評分不一致的音檔，並召開第二次評分會議，此會議主要目的為，再次說明各題型評分原則內容、各級分口語能力描述，以強化評分者對各級分標準之判讀，達成評分者評分共識。一般來說，第一次評分會議與第二次評分會議時間間隔約為五週。

而每次測驗評分工作結束後，針對評分結果較差之評分者，如偏嚴格、偏寬鬆或一致性表現較不理想，除透過上述第二次評分會議再訓練外；也個別提供各評分者其自身評分嚴格度及穩定性的統計分析結果，做為自我調整改善評分品質的參考依據，使評分者能更加掌握評分規準，避免未來再度出現評分過度嚴格或寬鬆的情況，改善其評分一致性。若持續觀察未見改善，則不予續聘。

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目標的潛在能力，通常可通過該測驗的信度與效度分析來進行整體性的評估。緣此，本節將討論 2016 年 11 月 6 日所舉辦的全國性流利精通級口語測驗正式考試之信、效度來總結本測驗之效能。

此次流利精通級口語測驗正式考試的實際到考人數為 66 人，但其中一位考生於考試中途離場而未能完成考試，故實際完成考試之有效受測者為 65 人。先由兩位評分者參與評分，以獨立閱卷的方式共同評閱有效受測者 65 名，共 455 筆答題音檔；再針對其中 22 筆給分不一致音檔，由兩位研發人員分別給分，每人評閱 11 筆音檔，進行第三人閱卷機制。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。此次測驗的信、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。換句話說，意即該測驗所獲得的測驗結果（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四類：第一，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；第二，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；第三，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；第四，評分者信度，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」遂成為影響受測者分數之主要因素。「評分者的評分一致性」大致可分為兩種類型：評分者間一致性（inter-rater consistency）及評分者內一致性（intra-rater consistency）。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）。

以下將詳述 2016 年流利精通級口語測驗正式考試之信度，以「評分者嚴格度變異」和「斯皮爾曼等級相關」（Spearman rank order coefficient）來評估評分者間一致性與評分者內一致性。

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，以及評分者內信度。由表 6 評分者嚴格度分析結果可知，以嚴格度平均值 0 作為標準來看，兩位評分者的嚴格度差異落在±0.5 logit 以內，分別為 0.023、-0.023，皆符合標準。

表 6 評分者嚴格度

評分者編號	評閱人數	觀察的平均值	嚴格度	標準誤	Infit MNSQ	Outfit MNSQ
C06	455	4.88	0.023	0.041	1.29	1.56
C05	455	4.99	-0.023	0.041	1.32	1.60

註：觀察的平均值表示評分者平均給分成績

(2) 斯皮爾曼等級相關

針對兩位評分者評分結果進行斯皮爾曼等級相關分析，以了解評分者間信度，結果如表 7 所示。兩位評分者各題平均值介於.973 至 1.000 之間，達到高度正相關，且均達.9 以上，評分者間信度相當良好。

表 7 評分者間斯皮爾曼等級相關

題號 / 相關係數	第一題	第二題	第三題	第四題	第五題	第六題	第七題
	1.000**	.975**	.975**	.992**	1.000**	.981**	.973**

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。結果如表 6 所示，由嚴格度標準誤（standard error；簡稱 S.E.）可發現，兩位評分者的標準誤皆為 0.041，顯示兩位評分者給分均具有自身的穩定性。再由 Infit MNSQ 及 Outfit MNSQ 介於 0.5 到

1.5 的標準，評估評分者自身給分一致性是否如模式所預期，結果顯示，兩位評分者的 Infit MNSQ 值皆在標準之內，而 Outfit MNSQ 數值則大於 1.5 的標準。

由於前者是加權的統計量，而後者是未加權的統計量，較容易受到個體差異大的資料影響，因此一般多以前者作為判斷資料是否符合測量模型的依據。是故，在 Infit MNSQ 與 Outfit MNSQ 結果有差異的情形下，主要參考 Infit MNSQ 的分析結果，整體而言，兩位評分者之評分者內一致性均佳，自身評分穩定性良好。

綜合上述結果，2016 年流利精通級口語測驗正式考試之評分者信度分析結果顯示，二位評分者嚴格度非常相近，差異在±0.5 logit 以內，斯皮爾曼等級相關也具有高度一致性，顯示評分者間信度良好；在評分者內一致性方面，二位評分者皆符合適配度標準，且自身變異小。

（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三大類：第一，內容效度（content validity），指的是測驗內容的相關證據；第二，建構效度（construct validity），即關於測驗架構的證據；第三，效標效度（criterion validity），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，在口語測驗中，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效度，可確保測驗相關內容皆是經由標準化程序而來，能作為內容效度的證據。通過測驗試題分析及因素分析，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。在受測者進行測驗時，收集其對自身口語能力的主觀評估，進行受測者自評結果與測驗結果之相關度分析，則屬同時效度，可做為效標效度的一種證據來源。

本測驗效度分析將分別由程序性效度（procedural validity）、試題分析、驗證

性因素分析（confirmatory factor analysis）以及受測者自評問卷與成績之相關分析等方面來說明 2016 年「華語文口語測驗」之內容效度、建構效度及效標效度。

1. 程序性效度

首先，本測驗研發人員在確立了評分方式和評分原則之後，針對評分者的培訓制訂了一套標準化流程，每次評分工作皆包含兩次評分會議。第一次會議的主要目的是調校評分者的評分標準，藉由讓評分者進行試評與討論，調整並統一評分者的評分標準；接著，再讓評分者獨立進行正式評分工作。正式評分工作結束後，便舉辦第二次評分會議；第二次會議的目的有二，一方面針對給分不一致的音檔進行討論，調整不一致的部分並建立評分共識；另一方面則是再次確定各級分之範例音檔，以強化評分者對各級分標準之判讀。第一次評分會議與第二次評分會議時間間隔約為五週。詳細評分流程參見表 8。

表 8　標準化評分流程

階段	工作項目	內容
1	第一次評分會議前置作業	研發人員從受測者答題音檔中挑選範例音檔做為第一次評分會議的試評音檔。
2	第一次評分會議	邀請評分者參與評分會議，現場進行試評工作，並依據試評結果面對面討論，建立評分共識。
3	正式評分	評分者透過線上評分系統各自進行為期二週的評分工作。
4	第二次評分會議前置作業	評分者透過線上評分系統繳交評分結果及評分依據，由研發人員加以整理，彙整出需要討論的音檔以及問題。
5	第二次評分會議	邀請評分者面對面討論，針對評分結果不一致的音檔，確立共識。
6	評分結果分析	將評分結果交由統計人員，分析評分者評分嚴格度、評分者間與評分者內一致性等資訊，作為未來評分培訓的參考。

透過標準化評分流程，流利精通級測驗之二位評分者嚴格度十分接近，嚴格度差異皆在±0.5 logit 以內，顯示所有評分者嚴格度符合標準，且皆具有自身評分一致性（詳見表 6），也就是說，各評分者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者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

(1) 試題分析

本測驗之組卷方式是依據試題反應理論而來。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對資料進行分析，因第一題曾用於其他次考試中，試題難度參數為已知，故固定難度參數來估計其他試題，結果如表 9 所示，第五題與第四題較難，難度參數分別為 3.993、3.837；第一題由於為銜接高階級與流利級的試題，故最為容易，難度為 0.822；第二題至第七題難度介於 2.802 至 3.993 之間，表示結果符合預期，難度較高，七道試題估計標準偏誤差異不大（介於 0.055 至 0.149 之間）。本測驗採用 Infit MNSQ 及 Outfit 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，試題與模式的適配情形皆良好，七道試題的數值都介於 0.5 至 1.5 之間，顯示本測驗試題測量到相同的潛在特質，也就是口語表達能力，意即流利精通級口語測驗正式考試具有一定程度的建構效度。

表 9 試題難度分布

試題編號	難度	標準誤	Infit MNSQ	Outfit MNSQ
第五題	3.993	0.149	1.29	1.15
第四題	3.837	0.137	0.96	0.97
第三題	3.591	0.140	0.70	0.70
第六題	3.268	0.144	0.84	0.90
第七題	3.013	0.139	0.92	0.91
第二題	2.802	0.142	0.90	1.04
第一題	0.822[A]	0.055	0.87	0.90

註： [A] 表示固定此題試題難度參數，不採自由估計。

再就試題鑑別度來看，一般二元計分題型使用點二系列相關係數作為試題鑑別度的指標，而口語測驗為多元計分題型，考生在題目的得分有多種不同情況，可改採皮爾森積差相關係數作為試題鑑別度指標。結果如表 10 所示，七道試題

相關係數介於.831 至.907 之間，一般來說，鑑別度.40 以上表示優良（郭生玉，2000），流利精通級七道試題的鑑別度均在.40 以上，表示鑑別度良好。

表 10　試題鑑別度分布

試題編號	積差相關係數
第一題	.907
第二題	.893
第三題	.907
第四題	.888
第五題	.831
第六題	.890
第七題	.877

(2) 驗證性因素分析

除了透過試題分析來評估本測驗是否具有建構效度之外，本報告亦從「驗證性因素分析」評估本測驗的建構效度。雖流利精通級測驗包含角色扮演及觀點回應類題型、文章摘要類與觀點論述類題型這三大類題型，欲分別測量「回應能力」、「摘要能力」與「論述能力」，但三者在測驗定義上，皆旨在測量口語表達能力，因此，為評估此三類測驗題型是否能夠組合為單維（uni-dimensionality）能力，即口語表達能力，本報告以結構方程模式（structural equation model）進行單因素模型驗證性因素分析（single factor model），以試題為測量變數，欲測得之能力為潛在變數；換言之，流利精通級測驗的測量變數為七道試題，潛在變數為口語表達能力。

在這部分的分析中，樣本為本次正式考試有效受測者共65人，此節使用Mplus 7.0版（Muthén & Muthén，2012）進行資料分析，估計方法採用「平均數與變異數修正後的加權最小平方值法」（weighted least squares means and variance adjusted；簡稱WLSMV），驗證性因素分析結果則分別透過基本適配度及整體適配度指標進行模式評估。

依據 Bagozzi 和 Yi（1988）以及 Hu 和 Bentler（1998），訂定出基本適配度的評估標準如下：（1）因素負荷量介於.50 至.95 之間；（2）相關係數不可大於1.0；（3）不能有過大的標準誤。至於整體適配指標，則採用卡方自由度比（χ^2/df）來評估整個模式與觀察資料的適配程度；以平方概似平方誤根係數（root mean

square error of approximation；簡稱 RMSEA）指標來評估整體模式的絕對適配度；以非規範適配指標（non-normed fit index；簡稱 NNFI，亦稱為 TLI）與比較適配指標（comparative-fit index；簡稱 CFI）二項指標來評估整體模式增值適配度。判斷標準分別為：χ^2/df <3、RMSEA <.08、CFI 和 NNFI >.90。

依照上述標準，在基本適配指標部分，流利精通級單因素模式分析結果（如圖 3），因素負荷量介於 .85 至 .95 之間，因素負荷量標準誤都達到顯著水準（p<.05）；所有數值均符合各項標準，表示單因素模式符合模型基本適配度之標準，顯示七道試題皆測得相同能力，即口語表達能力。經由初步檢驗，單因素驗證性因素分析模型適合解釋流利精通級口語測驗。

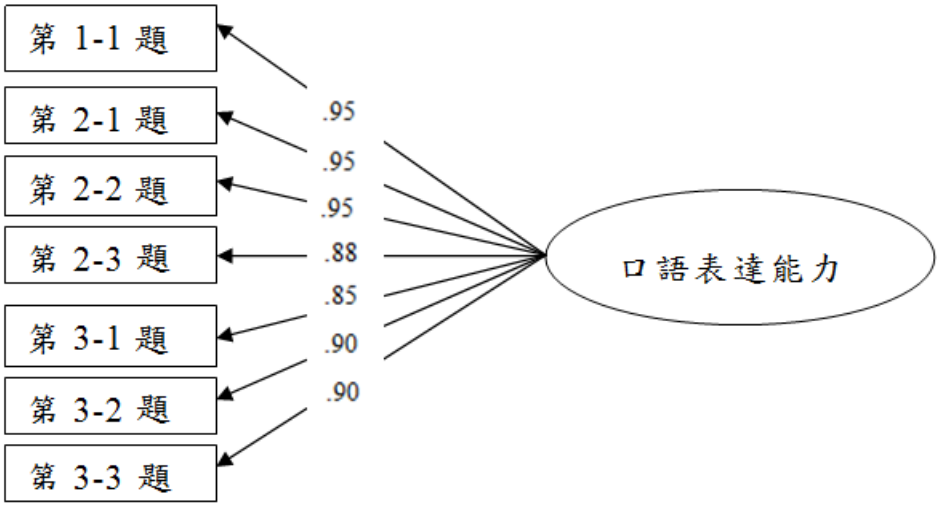

圖 3　流利精通級測驗單因素驗證性因素分析

整體模式適配度主要在評量整個模式與觀察資料的適合程度，相當於模式的外在品質。首先，經由卡方自由度比評估模式適配度，結果顯示 χ^2/df 值為 1.76，小於 3，表示模式適配度良好（Wheaton，Muthen，Alwin & Summers，1977）。

而在絕對適配度評估上，RMSEA 數值略高於.08，表示單因素模式不符合絕對適配度指標；在增值適配度評估部分，CFI 和 NNFI 數值皆大於.90，顯示單因素模式均符合增值適配度指標。

表 11　流利精通級測驗整體模式適配度摘要表

檢驗模型	卡方檢驗		絕對適配度	增值適配度	
	χ^2	χ^2/df	RMSEA	NNFI	CFI
單因素模式	24.651*	1.76	0.11	0.997	0.996

*表示 $p<.05$。

綜合上述結果可知，流利精通級口語測驗的單因素模式大致符合評估標準，具有建構效度，單因素模式可用以解釋測驗結果。

3. 效標效度

本測驗採用「受測者自評口語能力表現」做為效標來評估效標效度中的同時效度，以瞭解考生對於自己口語能力表現評估與實際測驗表現之間的關聯性。

於正式考試結束後，請受測者填答一份口語能力自評問卷（如附錄二所示）以收集相關資料進行同時效度分析，自評問卷採李克特五點量表（Likert scale），共有 8 道試題，受測者在閱讀完每道試題的能力描述後，從「總是可以」、「常常可以」、「有時可以」、「不常可以」及「很少可以」五個選項中，圈選出一個最符合的選項。計分方式為：圈選「總是可以」得 5 分；「常常可以」得 4 分；「有時可以」得 3 分；「不常可以」得 2 分；「很少可以」得 1 分。8 道試題回答結果之加總即為受測者口語能力自評結果，隨後再分別與其測驗總分進行相關分析。結果顯示，受測者自評結果與測驗總分的積差相關係數為.249（$p<.05$），顯示受測者自評口語能力與測驗總分間具有正相關存在，自評口語能力越佳者，其口語測驗總分越高。

從表 12 可知，自評問卷的 8 道問題中，相關係數達.05 顯著水準的題目為 Q2、Q3 和 Q5。以 Q2 和 Q3 而言，分別明確地要求考生就「單向觀點論述能力」、「彈性調整詞彙、語調以符合當時情境」這兩項能力自評，這可能讓受測者自評時較易客觀地答題，也與現行流利精通級題型有較直接的對應關係，可能因此使得自評結果與成績間的相關係數較高。至於 Q5 這一題的自評重點為能否「針對不同立場的看法，透過說明和評論的方式，進行回應、辯論並說服對方」，可能因該點為流利精通級這個階段最具代表性的口語表達能力項目，因此相關係數較高。

針對相關係數不顯著的題目，未來將持續追蹤，若相關係數數次皆不顯著，

將考量該題區辨性不足而予以刪除。

表 12　自評問卷各題與測驗總分之相關分析結果（N=65）

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8
測驗總分	.203	.275*	.268*	.228	.251*	.059	.110	.219

註：Q1-Q8 表示自評問卷題號；*表示 $p<.05$。

2016 年 11 月 6 日流利精通級口語測驗正式考試之效度指標顯示，透過標準程序訓練的兩位評分者，在評分上均具有一定穩定度，確保了一定程度的內容效度。所有試題皆測量到相同能力，具有建構效度；單因素驗證性因素分析結果也符合適配度指標，具有建構效度。受測者自評口語能力與測驗總分具有正相關存在，顯示測驗具有效標效度。

五、 結論

本文為 2016 年華語文口語測驗技術報告，闡述內容主要著重兩個部分，第一部分為針對流利精通級口語測驗之口語能力描述、測驗題型題數、評分規準及通過門檻等方面進行概述，並說明本測驗之整體性測驗研發、施測和評分之標準化流程。第二部分則主要針對於 2016 年度首次推出正式考試的流利精通級口語測驗之整體性測驗信度與效度評估，目的在檢視其是否能夠發揮測驗效用，確切地測量受測者的目標潛在口語能力。

在測驗信度分析方面，由於本測驗之受測者成績主要仰賴評分者判定，因此，受測者成績除了受到受測者自身具備之口語能力與測驗試題難度的影響之外，亦會受到評分者嚴格度變異的影響，故本測驗主要以「評分者自身給分穩定性」與「評分者間給分一致性」二個面向評估測驗信度。

在測驗效度分析部分，為使評分者皆能遵守測驗所擬定之評分原則，並據此給予受測者適切的評分，本測驗採取了標準化的評分流程來培訓評分者。此標準化流程為程序效度，可確保測驗相關內容皆經由標準化程序而來，為本測驗提供內容效度方面的證據。除了具備測驗之內容效度方面的證據之外，在施測完成後，本會也針對測驗所得之受測者作答反應資料，分別進行了試題分析與驗證性因素分析，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，是否與口語測驗研發之初所制訂的目標相同，並以此作為測驗之建構效度證據。最後，我們還透過受測者自評結果與受測者實際測驗結果的對照，來評估測驗結果的預測力，可以說，具有測驗之效標效度證據。

總體而言，從 2016 年度全國性流利精通級口語測驗正式考試之信度、效度分析的資料來看，可大致總結三項要點如下：

第一、所有評分者經由標準化評分訓練流程後，皆可達到評分者自身評分穩定性及評分者間評分一致性。換句話說，評分者可更好地掌握測驗之評分原則，並給予受測者適切的評分。

第二、受測者獲得的測驗成績與測驗研發之初所訂定之目標口語能力相符。

第三、受測者整體自評結果可作為測驗結果的有效預測效標。

綜上所述，2016 年度流利精通級口語測驗，其受測者成績具有可信度，可測得受測者之目標口語能力。

六、 文獻

高丙梁（2007）。计算机口试与面试的比较研究。**外语电化教学**。2007 年 4 月。第 114 期，73-76。

郭生玉（2000）。**心理與教育測驗**。台北：精華書局。

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。

蔡基剛（2005）。大学英语四、六级计算机口语测试效度、信度和可操作性研究。**外语界**。2005 年第 4 期（總第 108 期），66-75。

張蓉（2008）。大规模英语口语测试的思考。**浙江万里学院学报**。第 3 期 。136-138。

張蓉（2010）。试析 COPI 在我国大规模英语口语测试的应用前景。**內蒙古民族大學學報**。第 36 卷第 2 期 。100-102。

張濤（2009）。現行英語口語測試優劣分析。**天津职业院校联合学报**。第 3 期。80-82。

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（3）口語測驗信效度**（編號：ISBN 978-986-92167-4-6）。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（3）口語測驗信效度**（編號：ISBN 978-986-92167-9-1）。新北市：國家華語測驗推動工作委員會。

Bachman, L. F., & Savignon, S. J.（1986）. The Evaluation of Communicative Language Proficiency: A Critique of the ACTFL Oral Interview. *The Modern Language Journal*, 70（4）, 380-390.

Bagozzi, R.P., & Yi, Y.（1988）. On the evaluation of structural equation models. *Journal of the Academy of Marketing Science, 16*（1）, 74-94.

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

Clark, J.L.D., & Li, Y.C.（1986）. *Development, validation, and dissemination of a proficiency-based test of speaking ability in Chinese and an associated*

assessment model for other less commonly taught languages. Washington, DC: Center for Applied Linguistics.

Council of Europe.（2001）. *Common European Framework of Reference for Languages: learning, teaching, assessment* （chap.1 & chap.4）. Retrieved January 17, 2007, from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Hu, L.T., & Bentler, P.M. （1998）. Fit indices in covariance structure modeling: sensitivity to underparameterized model misspecification. *Psychological Methods, 3*, 424 – 453.

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.

Kenyon, D., & Tschirner, E. （2000）. The Rating of Direct and Semi-Direct Oral Proficiency Interviews: Comparing Performance at Lower Proficiency Levels. *The Modern Language Journal*, 84（1）, 85-101. Retrieved from http://www.jstor.org/stable/330451

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

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

Messick, S.（1996）.Validity of performance assessments. In G. W. Phillips (Ed.), *Technical issues in large-scale performance assessment*. Washington, DC: U.S. Department of Education, National Center for Education Statistics, 1-18.

Muthén, L.K. and Muthén, B.O.（2012）. Mplus（Version 7.0）[Computer Software]. Los Angeles, CA: Muthén & Muthén.

Nuessel, F.（1991）.Foreign Language Testing Today: Issues in Language Program

Direction.*Assessing Foreign Language Proficiency of Undergraduates.* Ed. T. V. Teschner. Boston, MA: Heinle and Heinle

Stansfield, C. W.（1989）.Simulated Oral Proficiency Interview. ERIC Digest. Washington. DC：Center for Aplied Linguistics.

Thrasher, R.（2000）.The testing of listening comprehension. Available Internet http://subsite.icu.ac.jp/people/randy/Lesson%20Six%20Text.pdf. 2000.

Wheaton, B., Muthen, B., Alwin, D., & Summers, G.（1977）. Assessing the reliability and stability in panel models. In D.R. Heise（ed.）, *Sociological Methodology*. San Francisco: Jossey-Bass.

附錄一　流利精通級口語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
1.會議帶領者對於本次會議的目的/任務解釋得很清楚。	3.92	98%
2.會議帶領者對於標準設定方法的操作流程說明得很清楚。	3.92	98%
3.我了解最低能力者在標準設定方法上的涵義。	3.83	96%
4.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。	3.92	98%
5.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。	3.92	98%
6.在第二回合，提供考生音檔級分有助於我判斷通過門檻分數。	4.00	100%
7.我是根據 C1 最低能力描述判斷 C1 通過門檻分數。（程序 3）	4.00	100%
8.我是根據 C2 最低能力描述判斷 C2 通過門檻分數。（程序 3）	3.92	98%
9.我對於自己所設定的通過門檻分數（cut score）有信心。	3.92	98%

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。

附錄二　華語文口語能力問卷-流利精通級

座位號碼：

考生姓名：

1.你填寫的資料只提供研究使用，填答結果**絕對保密**，**也絕對不會影響口語測驗成績**，請放心填寫。

You can be assured that the information you provide will be used ONLY for the academic purpose and will be completely confidential. In addition, it will not affect your test results.

2.請你想一想自己的口語能力，是不是能做到問題描述的內容，例如：「我能用中文寫信」，如果你「總是可以」做到「我能用中文寫信」，就在適當的圈內　塗黑「●」。

On a scale from 1 to 5 （with 1 being **rarely**, 2 **not often**, 3 **sometimes**, 4 **often**, and 5 **always**）, please indicate your opinion on the following statements. For example, if you feel "I can **always** do it" about the statement —"I can write letters in Chinese."— please fill in 5 as ●.

3.如果需要修改，請用橡皮擦修改，不要用修正液，請保持乾淨。

To make corrections, please use an eraser instead of white-out and keep the sheet clean.

※範例 Example：正確 Acceptable → ●　　不正確 Unacceptable → ⊙ ⃠ ⊗

請開始作答

Please begin answering the questions below.

		很少可以	不常可以	有時可以	常常可以	總是可以
1	無論是專業領域，或是具爭議性的議題，我都能詳細地說明自己的看法，並提出評論或建議。	1	2	3	4	5
2	在正式場合中，我能提出一份主題複雜的報告，透過清晰、有組織地說明及適當的例子，準確地表達自己的觀點。	1	2	3	4	5
3	不論對方使用什麼說話方式，我都能彈性地調整詞彙、語調，以符合當時的情境。	1	2	3	4	5
4	討論抽象、複雜的議題時，我能輕鬆地參與討論，並精確、流利地說明自己的觀點和反對其他立場的理由。	1	2	3	4	5
5	針對跟我立場不同的看法或批評，我能透過說明和評論的方式，輕鬆且正確地回應或辯論，並說服對方接受自己的立場。	1	2	3	4	5
6	遇到抽象、複雜或不熟悉的主題時，我能整合目前已知的資訊，做出清晰、有組織的報告，強調重點及相關細節，幫助聽眾注意重點。	1	2	3	4	5
7	針對相關主題，我能整合不同來源的資料和論點，進行全面性概述，總結意見，並以適當的方式結束發言。	1	2	3	4	5
8	在正式場合中發表言論時，我能靈活、精確地運用成語。	1	2	3	4	5

謝謝您的協助！Thank you very much for your help !

書　　名：華語文能力測驗技術報告—2016 (1)
口語測驗信效度
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