華語文能力測驗技術報告—2013（3）

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，「華語文口語測驗」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]，而每一等級又可再依據測驗成績細分為兩級，依序為入門級、基礎級、進階級、高階級、流利級、精通級，共六級。此架構相較於僅能區分受試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。

本報告分為三部分，首先將針對 2013 年新版口語測驗的內容、測驗實施與成績公布之標準化流程進行概述；其次，闡述本年度正式考試之信、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。

[1]流利精通級測驗尚在研發中。

二、 簡介

2013 年度本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），分別對應至歐洲共同語文參考架構（CEFR）之 B1（Threshold）與 B2（Vantage）。以下針對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）及通過門檻等方面，進行介紹。

（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，B1 等級學習者能針對個人感興趣的一或多個不同主題，使用線性順序的方式，進行直接的描述；B2 等級學習者則能進一步針對和興趣相關的廣泛主題，做出清晰且具系統性結構的描述，並說明相關細節，也能使用適當的例子和相關說明來支持自己的看法。而在互動能力的部分，B1 等級學習者能運用大量的簡單語言，針對熟悉的日常與非日常話題、感興趣或與職業相關領域的主題，進行資訊的交換、查證或確認，也開始能就抽象性的文化主題表達簡單的看法；B2 等級學習者則是能針對眾多不同的一般、學術、職業或閒暇話題，流利、自如且精確地表達看法，並能透過相關論證和舉例來強調事件與經驗對個人的重要性，以支持個人的看法。

本測驗綜合上述 CEFR 針對 B1、B2 等級所提出的口語能力描述，訂出進階高階級口語測驗各等級通過者所應具備的基本口語能力，如表 1 所示。

表 1　基本能力描述

通過等級	能力描述
進階級	1. 能直接且連貫地描述個人相關經驗及感覺、夢想、希望、真實或想像事件。 2. 能有次序地說明計畫或事件；能提出簡短的理由支持自己的看法。
高階級	1. 能清楚、仔細地描述感興趣的話題、經驗或事件。 2. 對於一般性議題或有爭議的內容，能提出個人見解、並有組織地詳細說明理由。 3. 能發展清晰的論點，舉出相關的例子延伸並支持自己的論點。

（二） 測驗題型及題數

擬訂出進階級、高階級的口語基本能力描述（如表 1）後，本測驗研發人員（以下簡稱研發人員）即循此方向設計測驗題型。

由於自進階級開始，學習者的描述能力已從入門基礎級時只能「使用簡單短語或句子，進行成串但不連貫地敘述」發展至「能直接且有組織性地描述」；此外，論證說明能力也開始萌芽發展，但入門基礎級階段的語言學習者尚不具備此類能力；因此大致可將此階段語言學習者的口語表達能力分為「描述能力」和「說明能力」二大面向。其中，進階級與高階級的第一條基本能力描述皆指向「描述能力」，而進階級的第二條基本能力描述，以及高階級的第二、三條基本能力描述則指向「說明能力」。因而，在進階高階級題型架構的設計上，針對「描述能力」和「說明能力」分別規劃了描述類和說明類兩大類題型；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。

描述類的題型依據試題內容的素材，再細分為「經驗描述題」和「圖片描述題」，前者著重於評量受測者能否「直接連貫且清楚地描述經驗與表達情感」，後者則透過受測者是否能客觀地描述圖片上的事件，並與個人經驗連結對照，評量受測者能否「直接連貫地描述真實或想像的事件」。說明類的題型著重於評量受測者能否「根據試題所提供的資料提出自己的意見和想法，並提出理由支持自己的論點」，以及「有組織地發展清晰的論點，舉出相關的例子」；據此，將說明類題型的內容設計為讓受測者做出選擇、提出建議、對議題表達贊成或反對立場等口語任務的「陳述意見題」。

而在受測者正式答題之前，為了讓受測者熟悉測驗方式，另設計了二題不計分的熱身題。進階高階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
進階高階級	熱身題	2
	經驗描述	2
	圖片描述	1
	陳述意見	3

另外，在作答時間的制訂過程中，本測驗參考了美國 AP 中文考試、中國 HSK 漢語水平考試、臺灣 GEPT 全民英檢及法國 DELF 法語鑑定文憑等語言能力測驗對於準備時間與回答時間的規定，並由本會所舉辦的全國性口語能力測驗預試中，分析受測者在各種測驗題型的回答時間，最後制定出進階高階級測驗的描述類題型，每一題的作答時間皆為 1 分 40 秒，說明類題型的每一題作答時間皆為 2 分鐘。

（三） 評分規準（rubric）

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。制訂評分規準（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；「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停頓次數、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。

其中，「內容組織」向度的考察重點包含任務完成度，而與口語任務的類別有關，反映的就是受測者的描述性能力與說明性能力；以描述類型的語言任務為例，通過進階級的語言使用者需「能直接且連貫地描述個人相關經驗及感覺、夢想、希望、真實或想像事件」，通過高階級的使用者需能「清楚、仔細地描述感興趣的話題、經驗或事件」。至於，說明類型的語言任務，通過進階級的語言使用者需「能表達尚稱清楚的論點，並提出尚能支持自己論點的簡短說明」，通過高階級的語言使用者需「能提出清楚且前後連貫的論點，且能夠透過相關例證延伸並支持自己的論點」。因而，描述類題型的「內容組織」向度著眼於「回答內容是否能直接描述，甚至達到清楚、詳述的程度」，說明類題型的「內容組織」向度則著眼於「論點是否清楚、相關說明或舉例是否能支持自己的論點」。因此描述類和說明類這兩大類題型無法共用「內容組織」向度的評分要點，而必須各自規劃對應的內容。

「語言運用」和「表達能力」用以評量的特質為語速、語音、詞彙和語法，這類語言特質為語言學習進程的共性，不受口語任務類別影響，因此可共用於描

述類和說明類這兩大類題型。

據此，研發人員針對描述類和說明類這兩大測驗題型各自規劃了一套對應的評分規準，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，根據各等級的基本口語能力指標、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、語法和語調）等方面，共同制定出進階高階級測驗描述類和說明類題型評分規準的內容，如表 3、表 4 所示，描述題評分原則適用於進階高階級的「經驗描述」及「圖片描述」題型，而說明題評分原則適用於「陳述意見」題型。而這兩種評分原則，僅在「內容組織」此向度的內容上有所差異，「語言運用」和「表達能力」這兩向度的內容則為相同。

本測驗採整體式評分，描述題與說明題評分級距皆設定為 0 至 5 級分，評分教師聆聽錄音檔案後，依據評分原則三大向度描述內容，給予一整體分數。

表 3　進階高階級描述題評分原則

級分	內容組織	表達能力	語言運用
5	回答內容已能完成題目要求、豐富，描述清楚、詳細；話語有組織、前後連貫。	表達順暢、流利，少有停頓；語音正確、清楚，都能被聽者理解。	具備多樣的詞彙、語法結構，能適當、較準確地使用，僅有少許錯誤。
4	回答內容已能完成題目要求、尚稱豐富，描述大致清楚、詳細；話語多有組織、前後大致連貫。	表達尚稱順暢、尚稱流利，仍有一些停頓；語音大致正確、大致清楚，幾乎都能被聽者理解。	具備足夠的詞彙、語法結構，幾乎都能正確、適當使用，仍有一些限制、錯誤。
3	回答內容已能完成題目要求，內容尚稱充足，已能進一步加以描述，話語尚有組織。	語速適中，偶有停頓；語音大致清楚，偶有錯誤，大致都能被聽者理解。	具備足夠的詞彙和語法結構，能大致適當使用，偶有錯誤。
2	回答內容已能完成題目要求，內容稍嫌不足，以至於有時無法進一步加以描述，部分話語缺乏組織。	語速稍慢，常有停頓；詞語重複次數多；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，有時無法直接表達較複雜的意思，時有錯誤。
1	回答內容已完成題目要求，但內容不足，話題無法擴展，組織較差。	語速緩慢，說話費力，停頓次數過多且時間過長；詞語重複次數過多；語音多不正確，聽者較難理解。	大致掌握基本詞彙和簡單的語法結構，且達到基本溝通需求，仍偶爾犯基本、規律的錯誤。
0	考生靜默，沒回答；離題。		

表 4　進階高階級說明題評分原則

級分	內容組織	表達能力	語言運用
5	論點清楚、前後連貫；提出的理由都能支持自己的論點，內容豐富，話語有組織。	表達順暢、流利，少有停頓；語音正確、清楚，都能被聽者理解。	具備多樣的詞彙、語法結構，能適當、較準確地使用，僅有少許錯誤。
4	論點清楚、前後連貫；提出的理由幾乎都能支持自己的論點；少有重複說明的情況；內容尚稱豐富，話語多有組織。	表達尚稱順暢、尚稱流利，仍有一些停頓；語音大致正確、大致清楚，幾乎都能被聽者理解。	具備足夠的詞彙、語法結構，幾乎都能正確、適當使用，仍有一些限制、錯誤。
3	論點大致清楚；提出的理由大致能支持自己的論點，偶有重複說明的情況，但已能進一步解釋，內容尚稱充足，話語尚有組織。	語速適中，偶有停頓；語音大致清楚，偶有錯誤，大致都能被聽者理解。	具備足夠的詞彙和語法結構，能大致適當使用，偶有錯誤。
2	論點稍嫌不清楚；提出的理由尚能支持自己的論點，仍常有重複說明的情況，以至於有時無法進一步加以解釋，內容稍嫌不足，組織稍差。	語速稍慢，常有停頓；詞語重複次數多；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，有時無法直接表達較複雜的意思，時有錯誤。
1	論點不甚清楚；提出的理由不足以支持自己的論點；內容不足以完成題目要求，組織較差。	語速緩慢，說話費力，停頓次數過多且時間過長；詞語重複次數過多；語音多不正確，聽者較難理解。	大致掌握基本詞彙和簡單的語法結構，且達到基本溝通需求，仍偶爾犯基本、規律的錯誤。
0	考生靜默，沒回答；離題。		

（四） 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。由於進階高階級測驗給分方式為 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，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，詳細內容可參考藍珮君、陳柏熹、張可家、施泰亨和林玲英（2013）。

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. 說明描述題題型內容與評分原則，播放各級分範例音檔，藉由考生實際的答題反應，具體化評分原則。
4. 請成員依據提供的進階級、高階級最低能力描述，分別與描述題的評分原則進行配對，決定進階級和高階級口語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. 請成員聽完描述題題型的 10 個應試者音檔後，依據進階級、高階級最低能力描述，分別判斷每個音檔的 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)每個音檔被判定為 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音檔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音檔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描述題之通過門檻。

設定進階高階級說明題之通過門檻時，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.每一回合音檔 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，進階級通過門檻部分，描述題的標準差在第一回合最大，然後逐漸降低，說明題可能經過上午描述題的討論後，專家們判斷原則漸趨一致，在第二回合 14 位專家的判斷已達到完全一致，標準差為 0，高階級通過門檻的標準差也呈現同樣的情形。

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	描述題	0.535	0.363	0.267
	說明題	0.267	0.000	0.000
高階級	描述題	0.646	0.363	0.000
	說明題	0.267	0.000	0.000

描述題與說明題各 10 個音檔 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，與音檔實際級分求相關的結果，14 位專家描述題三個回合的相關係數均介於.703 至.949 之間(p<.01)，說明題三個回合的相關係數依序為.730 至.949 (p<.01)、.791 至.949 (p<.01)、.783 至.949 (p<.01)，顯示專家們對於音檔通過等級的判斷與實際得分之間具有中高度或高度的正相關存在，判斷結果與實際得分頗為一致。

華語文口語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口語測驗，能有效將華語學習者的口語表現區分為 CEFR 的 B1 和 B2 兩等級。

根據標準設定研究結果，本測驗進階高階級的計分題目共有六題，各題均採

0 至 5 級分的評分級距，考生測驗總分為六題計分題的成績加總，滿分為 30 分，各等級通過分數範圍如表 6 所示。測驗總分介於 12 至 23 分者，可取得進階級（Level3）證書，總分介於 24 至 30 分者，可取得高階級（Level4）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	24-30
	進階級	12-23

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。

2013 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：

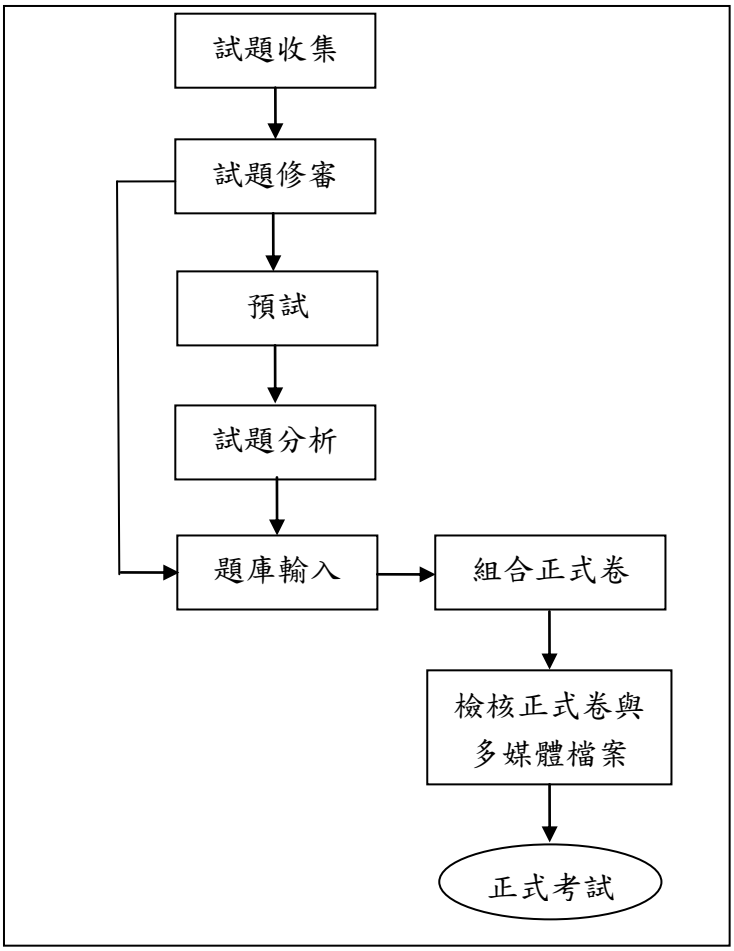

圖 1　正式考試製卷流程

1.　試題收集

2013 年度本測驗命題教師共計八位（含入門基礎級、進階高階級），每位教師每期繳交一套試題，每期命題時長約兩個月。命題教師在正式命題前，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、口語基本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教師命題指導文件及《華語八千詞詞表》[2]。本年度的命題回收數量為：入門基礎級 78 題、進階高階級 62 題，共計 140 題。

2.　試題修審

新版口語測驗將入門級、基礎級合併為入門基礎級，進階級、高階級合併為進階高階級，由於各等試卷涵蓋兩個等級，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

[2]《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.rar

能力。

(1) 會內初審

待命題教師繳交試題後，即由研發人員進行第一階段初審工作，隨後回覆審題意見，命題教師再根據審題意見修改試題，修改時間約為二週。

(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，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。本會分別於 2013 年 3 月及 7 月舉辦口語測驗進階高階級、入門基礎級全國預試。到考人數分別為：入門基礎級 102 名、進階高階級 87 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教師人工判定（詳見 P.15 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之多相模式（facets model）（Linacre，1989），對預試資料進行分析。由於計分採級分制（入門基礎級為 0-3 級分、進階高階級為 0-5 級分），屬多元計分方式的試題，因此，本測驗使用可進行多相模式分析之 Facets 3.71.3 版之部分給分模式（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 Facets3.71.3 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 InfitMNSQ 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。

2013 年兩次預試包括入門基礎級九道試題、進階高階級六道試題，試題分析結果顯示，所有試題之 Infit 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。本年度輸入口語測驗題庫試題數量總計為 64 題，分別為入門基礎級 40 題、進階高階級 24 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。本會於 2013 年 11 月 3 日舉辦華語文口語測驗進階高階級正式考試，題型與題數如表 2 所示。

7. 檢核正式卷與多媒體檔案

研發人員需於每次施測前兩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：

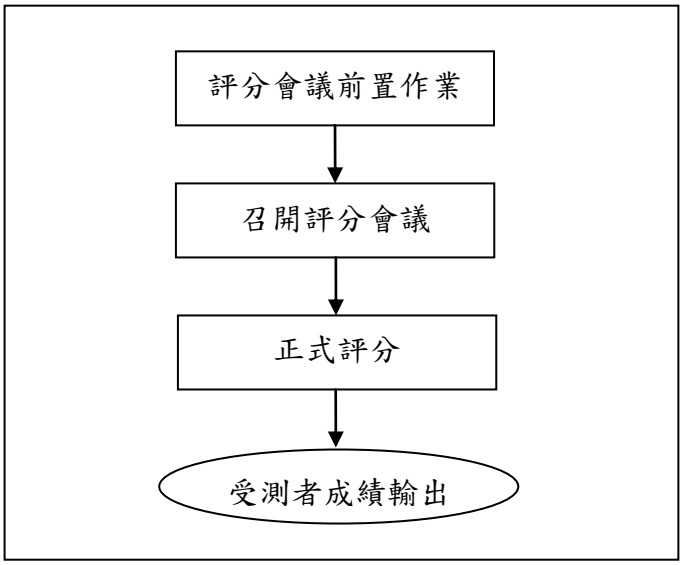

圖 2　評分流程

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。各等級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。評分教師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。2013 年度召開的第一次口語測驗正式考試評分會議，共有八位評分教師參加。

3.　正式評分

評分會議後，評分教師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教師進行評分，評分方式採整體式評分（holistic scoring），評分時間約為兩週。

4. 受測者成績輸出

在評分教師完成正式評分並繳交當次評分結果及其評分依據說明後，本會即彙整各評分教師之評分結果與依據，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，有兩名以上評分教師評定級分不一致，即交由第三位評分教師評定成績。採用此給分方式，較能避免「針對同一名受測者，不同評分教師評分結果差異過大」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，有兩名以上評分教師評分不一致音檔，召開第二次評分會議，此會議主要目的為，再次說明評分規準各級分口語能力描述，以強化評分教師對各級分標準之判讀，達成評分教師在評分上共識，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。

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 2013 年 11 月 3 日所舉辦的全國性進階高階級口語測驗正式考試之信、效度來總結本測驗之效能。

此次正式考試的報名人數為 151 人，實際到考人數為 139 人。由九位評分人員參與評分，分別評閱 59 至 60 名受測者之答題音檔。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 2013 年華語文進階高階級口語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）來評估評分者間信度。

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，檢視評分教師嚴格度差異，以及評分者內信度。由表 7 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，九位評分者中有五位評分嚴格度相差在±0.3 logit 以內，評分嚴格度相當一致；有二位評分者的嚴格度差異落在±0.3~±0.5 logit 範圍內，較他人差異略大；其餘二位評分者則超出±0.5 logit 之外，嚴格度與他人較不一致。其中，以編號 B04 教師給分較為嚴格，嚴格度為 0.734，B03 較為寬鬆，嚴格度為-0.526。

在評分教師自身給分穩定性上，由嚴格度標準誤（S.E.）可發現，各評分者的標準誤（S.E.）介於 0.076 至 0.078 之間，顯示出評分者給分均具有自身的穩定性，且各評分者的穩定性大致相當。再由統計指標（Infit MNSQ）介於 0.5 到 1.5 的標準，評估評分教師自身給分一致性是否如模式所預期，結果顯示，所有評分教師之評分者內一致性均佳，自身評分穩定性良好。

表 7　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（standard error）	Infit MNSQ
B04	60	2.34	0.734	0.078	0.96
B08	59	2.68	0.398	0.078	1.15
B07	59	2.79	0.169	0.078	1.27
B15	60	2.89	0.002	0.076	0.94
B16	139	2.87	-0.058	0.050	0.97
B06	59	2.92	-0.102	0.078	0.99
B05	60	2.78	-0.208	0.078	1.06
B14	60	3.04	-0.307	0.076	1.49
B03	60	3.15	-0.526	0.076	0.85

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均方差；B16 為本會研發人員，評閱所有受測者音檔。

2. 評分者間信度

針對三組共九位評分者評分結果進行斯皮爾曼等級相關分析，以了解各組兩兩評分者的評分者間信度，結果如表 8 所示。評分者 B03、B14、B15 與 B16 這組，可能因 B03 評分較為寬鬆，故各題評分者間信度之平均值略低，均在.8 以

下，但仍達到.01 之顯著水準。評分者 B07、B08、B06 與 B16 評分者間信度良好，六道試題相關係數平均值皆達到.8 以上，具有高度正相關。而 B04、B05 與 B16 一組，各題平均值介於.743 至.849 之間，達到中高度至高度正相關，評分者間信度大致良好。

表 8　評分者間斯皮爾曼等級相關

題號 / 組別	第一題	第二題	第三題	第四題	第五題	第六題
B03 B14 B15 B16	.600	.713	.689	.728	.675	.675
B07 B08 B06 B16	.829	.821	.830	.878	.897	.862
B04 B05 B16	.765	.821	.773	.743	.830	.849

註：細格內數值為兩兩評分者間相關係數之平均值。

由上述可知，2013 年進階高階級口語測驗正式考試之評分者信度顯示，評分教師中，評分者 B07、B15、B16、B06 與 B05 評分嚴格度相差較小，一致性較高；而 B04 給分較為嚴格，B03 則較為寬鬆。在評分者內一致性方面，所有評分教師皆符合適配度標準，且自身變異未過大。在評分者間一致性方面，則是大部分教師評分者間信度大致良好，斯皮爾曼等級相關具有中高度至高度的一致性；有一組評分者間信度較不理想。

為了確保評分教師的評分品質，針對評分結果較差，如偏嚴格、偏寬鬆之評分教師，本會透過第二次評分會議進行評分再訓練（如表 9）。此外，也個別提供各評分教師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，使評分教師能更加掌握評分規準，給予受測者更為客觀、合理、適切的成績，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。同時也會將這些評分教師列入觀察名單，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」（performance assessment），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、因素分析之驗證性因素分析（confirmatory factor analysis）及同時效度（concurrent validity）等四方面來描述 2013 年「華語文口語測驗」之內容效度、建構效度及效標關聯效度。

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，針對評分教師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，藉由讓評分教師進行試評與討論的辦法，來調整並統一評分者的評分標準；接著，再讓評分教師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，以強化評分教師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 9。

表 9　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	邀請評分教師參與評分會議，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	評分老師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	評分老師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	邀請評分教師面對面討論，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，分析評分教師評分嚴格度、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，進階高階級測驗之九位評分教師嚴格度雖然有所不同，如評分教師 B04 嚴格度偏嚴，B03 則偏鬆；然而，九位評分者中有七名評分者嚴格度相差±0.5 logit 以內，顯示多數的評分者嚴格度相近，且所有評分教師皆具有自身評分一致性（詳見表 7），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

(1)　試題分析

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 10 所示，第六題、第四題與第一題較難，難度參數分別為 0.369、0.289 以及 0.246；第五題、第二題與第三題較容易，難度依序為-0.122、-0.104 以及-0.051；六道試題估計標準誤差異不大（介於 0.05 之間）。本測驗又採用 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），結果如表 10，試題與模式的適配情形皆良好，六道試題的 Infit MNSQ 值都介於 0.7 至 1.3 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，意即本測驗進階高階級正式考試具有一定程度的建構效度。

此外，從各題試題難度參數的分布情形，也可了解測驗結果與命題預期難度是否相符。試題編號第一題至第三題為描述類題型，預期難度應較第四至第六題說明類題型容易，各題難度分布大致呈現此一趨勢，惟第一題稍難而第五題較為容易。

檢視試題內容，研發人員推測可能由於第五題主題與網路使用有關，與受測者生活經驗貼近，且在說明時不一定要使用較困難的詞彙也能完整表達，故難度較低；而第一題可能因此一問題的問法未限定受測者描述當時的情形，受測者若於回答時描述的比重較低，說明的比重較高，評分時會產生歧異，不易得到高分，故難度略難。未來命題時會更明確說明受測者要回答的內容，並於評分時與評分教師確認評分的判斷標準，以使測驗結果與預期難度更為一致。

表 10　試題難度分布

試題編號	難度	標準誤（S.E.）	Infit MNSQ
第六題	0.369	0.052	0.98
第四題	0.289	0.051	1.11
第一題	0.246	0.054	1.10
第三題	-0.051	0.053	0.94
第二題	-0.104	0.051	1.10
第五題	-0.122	0.052	0.84

註：Infit MNSQ 表示訊息加權適配度均方差。

(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。由於進階高階級測驗包含描述題與說明題這兩種題型，欲分別測量描述能力與說明能力，故主要以結構方程模式（structural equation model）進行二因素驗證性因素分析（correlated two factor model）。不過，進階高階級測驗雖包含兩種題型，但在測驗定義上，旨在測量口語表達能力，因

此，為評估此兩種測驗題型是否能夠組合為單維（uni-dimensionality）能力，即口語表達能力，本報告同時也進行了單因素模型驗證性因素分析（single factor model）。每種因素分析模型中，試題為測量變數，欲測得之能力為潛在變數。例如，單因素模型中，進階高階級測驗的測量變數為六道試題，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受測者共 139 人，使用 LISREL 8.51 版進行資料分析，其中，驗證性因素分析模型分別透過基本適配度及整體適配度指標進行模式評估。關於模式基本適配標準，根據 Bagozzi 和 Yi（1998）的研究，指標評估標準有四：

1. 誤差變異量不可為負；
2. 誤差變異量必須達到顯著水準；
3. 因素負荷量須介於.5~.95 之間；
4. 不能有很大的標準誤。

依照上述標準，在基本適配指標部分，進階高階級單因素模式分析結果（如圖 3），完全標準化因素負荷量介於 0.78 至 0.88 之間，標準誤介於 0.073 至 0.085 之間，誤差變異均達到顯著水準（$t_{.975,9}$=2.26）；完全標準化誤差介於 0.22 至 0.39 之間，標準誤在 0.042 至 0.072 之間，誤差變異同樣都達到顯著水準。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。而二因素模式分析結果（如圖 4），完全標準化因素負荷量介於 0.78 至 0.88 之間，標準誤在 0.074 至 0.085 之間，誤差變異均達到顯著水準（$t_{.975,8}$=2.31）；完全標準化誤差介於 0.23 至 0.39 之間，標準誤在 0.044 至 0.072 之間，誤差變異同樣也達到顯著水準。此外，描述能力與說明能力二潛在因素之間的完全標準化相關係數（φ）為 1.01，標準誤為 0.019；二因素模式所有數值均符合上述四項模式基本適配標準。綜上所述，兩種測驗題型能力具有高相關，顯示皆測得相同能力，即口語表達能力。經由初步檢驗發現，單因素與二因素驗證性因素分析模型皆適合解釋進階高階級口語測驗。

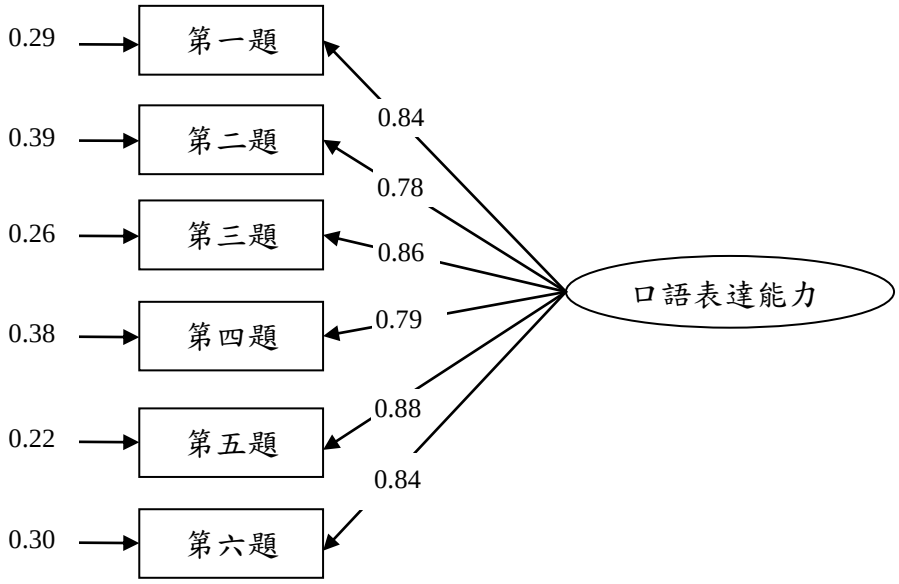

圖 3　進階高階級測驗單因素驗證性因素分析

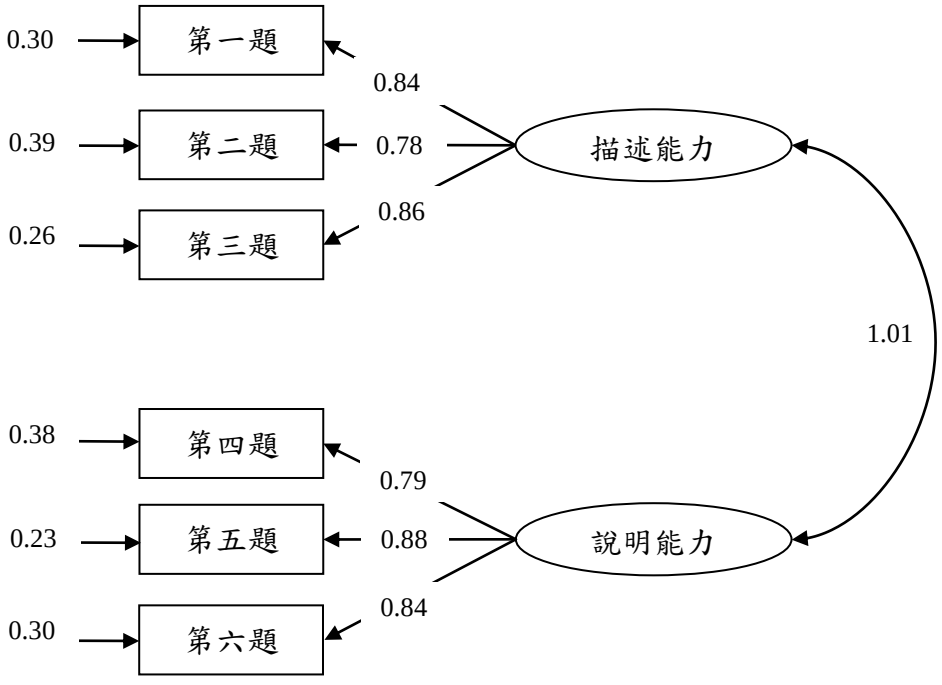

圖 4　進階高階級測驗二因素驗證性因素分析

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，經由卡方考驗發現，進階高階級單因素與二因素模式的 χ^2 值分別為 1.380（p=0.998）及 1.290（p=0.996），二種模式皆未達顯著水準，表示單因素與二因素模式的驗證模型均與觀察指標適配。而在整體適配度指標部分，

則根據適配指標(goodness-of-fit index；簡稱 GFI)、標準化殘差均方根指標(standardized root mean square residual；簡稱 SRMR)與平方概似平方誤根係數(root mean square error of approximation；簡稱 RMSEA)三項指標來評估整體模式的絕對適配度。以規範適配指標(normed fit index；簡稱 NFI)、非規範適配指標(non-normed fit index；簡稱 NNFI)與比較適配指標(comparative-fit index；簡稱 CFI)等三項指標來評估整體模式增值適配度。各適配指標如表 11 所示。

在絕對值適配度評估方面，各指標判斷標準分別為：GFI>0.90、SRMR<0.08 及 RMSEA<0.05。由下表可知，進階高階級測驗單因素與二因素模式指標皆符合前述適配度標準。而在增值適配度評估上，NFI、NNFI 及 CFI 三項指標的判斷標準均為>0.90，根據下表，兩種模式指標亦均符合標準。

表 11　進階高階級測驗整體模式適配度摘要表

檢驗模型	絕對適配度			增值適配度		
	GFI	SRMR	RMSEA	NFI	NNFI	CFI
單因素模式	0.997	0.007	0.000	0.998	1.021	1.000
二因素模式	0.997	0.007	0.000	0.998	1.021	1.000

綜合上述結果可知，進階高階級口語測驗無論是單因素或二因素模式，都符合評估標準，且指標數值相當接近，同樣都具有建構效度，二種模式都可用以解釋測驗結果。

3. 效標效度

本測驗採用兩種指標評估效標效度中的同時效度，分別是「受測者使用中文與家人交談頻率和實際測驗表現之間的關聯性」和「受測者自評口語能力表現與實際測驗表現之間的關聯性」。

於正式考試前，先請受測者填答使用中文與家人交談的頻率；並在正式考試結束後，請受測者填答一份口語能力自評問卷（如附件 2 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 17 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。17 道試題回答結果之加總即為受測

者口語能力自評結果[3]，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.294（p<.01），與測驗通過等級的等級相關係數為.202（p<.05），顯示受測者自評口語能力與測驗總分和通過等級之間均有正相關存在，自評口語能力越佳者，其口語測驗總分越高，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 12 所示。自評問卷中，有 12 題的答題反應與測驗總分的相關達到.05 或.01 的顯著水準，相關係數介於.195 至.273 之間，表示在這 12 題中回答可以做到頻率越高的受測者，其測驗總分也越高；其餘 5 題答題反應與測驗總分之間則沒有顯著關聯性存在。而與通過等級的相關，有 10 題相關係數達顯著水準（p<.05 或 p<.01），數值介於.171 至.244 之間，表示在這 10 題回答可以做到頻率越高的受測者，通過測驗等級也越高。

整體來看，相關係數較低的為 Q3、Q11 和 Q13，這三題的題目內容中分別以「完整」、「詳細」或「強調重點及細節」為自評的重點，可能內容較為抽象，考生判斷上較為困難，而造成相關係數較低。

另外，相關係數較高的題目為 Q5、Q 6 和 Q 7，這三題相較於其他題目而言，題目本身的敘述中較未使用任何彰顯程度高低的形容詞，這可能讓受測者自評時較易客觀地進行答題，使得自評結果與成績間的相關係數較高。

針對相關係數較低或不顯著的題目，未來將持續追蹤，若相關係數長期偏低，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、通過等級之相關分析結果

	Q1	Q2	Q3	Q4	Q5	Q6	Q7	Q8	Q9
測驗總分	.229**	.158	.104	.227**	.258**	.256**	.273**	.208*	.181
通過等級	.239**	.171*	.130	.136	.231**	.227**	.244**	.186*	.173
有效樣本	139	139	139	139	138	138	115	116	115
	Q10	Q11	Q12	Q13	Q14	Q15	Q16	Q17	
測驗總分	.195*	.152	.243**	.182	.195*	.218*	.238*	.239*	
通過等級	.120	.138	.223*	.133	.199*	.160	.191*	.205*	
有效樣本	116	115	116	116	116	114	116	116	

註：Q1-Q17 表示自評問卷題號；*表示 p＜.05；**表示 p＜.01。

[3]有 28 名考生未填答所有題目，問卷總分之有效樣本為 111 人。

表 13 為填答「使用不同頻率與家人交談」之受測者在進階高階級測驗中的成績表現，分為「經常」與「非經常」兩類別，回答經常與家人說中文的受測者，平均分數高於非經常的受測者。進一步以獨立樣本 *t* 檢定對此二種受測者進行測驗平均得分的考驗，結果發現，與家人用中文交談頻率不同的受測者，在測驗平均得分上達到顯著差異水準（$p<.01$），表示經常以中文與家人交談的受測者，在進階高階級口語測驗的表現優於非經常與家人用中文交談者。

表 13　經常與非經常與家人使用中文交談受測者之成績表

說中文頻率	人數	平均分數	*t* 值
（1）經常	26	20.0	2.925**
（2）非經常	113	16.4	

註：**表示 $p<.01$。

2013 年 11 月 3 日華語文口語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九位評分教師，在評分上均具有一定穩定度，確保了一定程度的內容效度。所有試題皆測量到相同能力，具有建構效度；單因素和二因素驗證性因素分析結果也都符合適配度指標，具有建構效度。受測者自評口語能力與測驗總分、通過等級皆有正相關存在；「經常」使用中文與家人交談的受測者，平均測驗成績優於「非經常」使用中文與家人交談的受測者，顯示測驗具有效標關聯效度。

五、 結論

本文為 2013 年華語文口語測驗技術報告，闡述內容主要著重兩個部分，第一部分為針對華語文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，並說明測驗研發、施測和評分之標準化流程。第二部分則主要進行 2013 年度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教師判定，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，亦會受到評分教師嚴格度變異的影響，故本測驗主要以「評分教師自身給分穩定性」與「評分教師間給分一致性」二個面向評估測驗信度。在此，要補充說明一點，本測驗於 2013 年開始，為了確保評分教師能切實掌握口語測驗之評分規準，給予受測者更為適切的評分，於每次正式評閱結束後，回饋評分教師評分嚴格度等相關訊息，協助其了解自身評分情形。此外，擬針對評分嚴格度變異較大或與評分一致性較低的評分教師，進行評分再訓練，並將這些評分教師列入觀察名單，若日後評分結果未獲改善，即不予續聘。

在測驗效度分析部分，為使評分教師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教師。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 2013 年度全國性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、所有評分教師經由標準化評分訓練流程後，幾乎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，評分教師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、受測者自評結果多數可作為測驗結果的有效預測效標。例如，自評口語能力較高或使用中文與家人交談頻率越高者，其測驗成績也較高。

綜上所述，本年度進階高階級口語測驗，其受測者成績具有可信度，可測得受測者之目標口語能力；測驗架構上也因一次涵蓋兩個等級，較過去的舊版測驗更能發揮測驗效能。

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。

藍珮君、陳柏熹、張可家、施泰亨、林玲英（2013 年 11 月）。**Yes/No Angoff 法在華語文口語測驗的應用**。2013 年第二屆標準本位評量國際研討會。臺北國立臺灣師範大學。

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.

Council of Europe.（2001）. *Common European Framework of Reference for Languages: learning, teaching, assessment* (chap.1 & chap.4). Retrieved January17, 2007, from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

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.

附件 1　進階高階級口語測驗標準設定研究問卷調查結果

問卷內容	平均數
1.會議帶領者對於本次會議的目的/任務解釋得很清楚。	4.0
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。	4.0
3.我了解最低能力者在標準設定方法上的涵義。	3.8
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。	3.8
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。	3.8
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。	3.6
7.我是根據 B1 最低能力描述判斷 B1 通過門檻分數。(程序 3)	3.8
8.我是根據 B2 最低能力描述判斷 B2 通過門檻分數。(程序 3)	3.8
9.我對於自己所設定的通過門檻分數(cut score)有信心。	3.8
10.我是根據 B1 最低能力描述判斷考生音檔 CEFR 等級。(程序 4)	3.8
11.我是根據 B2 最低能力描述判斷考生音檔 CEFR 等級。(程序 4)	3.8

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　進階高階級口語測驗問卷

座位號碼：

考生姓名：

1. 您所提供的資料僅供研究使用，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. 看完下面的問題以後，請你想一想自己的口語能力，是不是可以做到問題所描述的任務。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」，就圈選⑤。
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."— pleasecircle ⑤.

3. 如需修正，請以橡皮擦修改，勿使用立可白或其他修正液，並請保持乾淨。
To make corrections, please use an eraser instead of white-out and keep the sheet clean.

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	**對於學業或專業領域的話題，我能不費力做出前後連貫的描述。** I can give a reasonably fluent description of a subject within my academic or professional field, presenting it as a linear sequence of points	1	2	3	4	5
2	**對於我感興趣的或學習相關的領域，我能直接的描述我熟悉的話題。** I can give straightforward descriptions on a variety of familiar subjects related to my own fields of interest or study.	1	2	3	4	5
3	**我能完整的描述我的經驗、感覺和反應。** I can talk in detail about my experiences, feelings and reactions.	1	2	3	4	5
4	**我能描述一本書或一部電影的情節，並說出我的看法。** I can talk about the plot of a book or film and give my opinion.	1	2	3	4	5
5	**我能描述夢想、希望和抱負。** I can describe dreams, hopes and ambitions.	1	2	3	4	5
6	**我能敘述一個故事。** I can tell a story.	1	2	3	4	5

		很少可以	不常可以	有時可以	常常可以	總是可以
7	大部分的情況下，我提出的論點容易被理解。 I can develop an argument well enough to be followed without difficulty most of the time.	1	2	3	4	5
8	我能簡短的解釋和說明我的意見和計畫。 I can briefly explain and give reasons for my opinions and plans.	1	2	3	4	5
9	在預先準備的情況下，我能簡短的發表跟我日常生活領域有關的事項。 I can deliver short rehearsed announcements and statements on everyday matters within my field.	1	2	3	4	5
10	在預先準備的情況下，對於我熟悉的主題，我能做出簡單、清楚、易懂的口語報告，且報告中的重點能被理解。 I can give a simple, prepared presentation on a familiar topic within my field that is clear and precise enough to be followed without difficulty most of the time and in which the main points can be understood.	1	2	3	4	5
11	我能提出詳細的理由以支持自己的論點。 I can argue for my point of view in detail.	1	2	3	4	5
12	我能建立一個合理的、有邏輯的論點。 I can construct a chain of reasoned argument, linking my ideas logically.	1	2	3	4	5
13	我能做出清晰、有組織的敘述，並能強調重點及相關細節。 I can give a clear, systematically developed presentation, with highlighting of significant points and relevant supporting detail.	1	2	3	4	5
14	我能就一個問題提出自己的觀點，並說明不同選擇之下的優點和缺點。 I can explain a viewpoint on a topical issue giving the advantages and disadvantages of various options.	1	2	3	4	5
15	即使聽眾提出了我事前沒準備到的問題，我也能自然地回答。 I can depart spontaneously from a prepared text and follow up points raised by an audience.	1	2	3	4	5
16	我能發展清楚、有邏輯的論點，並使用適當的例子來延伸及支持自己的論點。 I can develop a clear coherent argument, linking ideas logically and expanding and supporting my points with appropriate examples.	1	2	3	4	5
17	我能對自己有興趣的相當廣泛的主題，清楚、詳細的描述，並使用適當的說明及例子來擴展和支持自己的論點。 I can give clear detailed descriptions on a wide range of subjects relating to my field, expanding and supporting ideas with subsidiary points and relevant examples.	1	2	3	4	5

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