華語文能力測驗技術報告—2014（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） 測驗題型及題數.....3
（三） 評分規準.....4
（四） 通過門檻.....6
三、 測驗標準化流程.....10
（一） 正式考試製卷流程.....10
（二） 評分流程.....14
四、 測驗評估.....16
（一） 信度.....16
1. 評分者內信度.....17
2. 評分者間信度.....17
3. 研究子題－評分者評分偏誤研究.....18
（二） 效度.....21
1. 程序性效度.....21
2. 建構效度.....23
3. 效標效度.....27
五、 結論.....29
六、 文獻.....31

表目錄

表 1　基本能力描述 3
表 2　測驗題型與題數分布 4
表 3　入門基礎級回答問題評分原則 6
表 4　入門基礎級描述題評分原則 6
表 5　標準設定各回合判斷結果之標準差 8
表 6　入門基礎級口語測驗通過門檻分數 9
表 7　評分者嚴格度 17
表 8　評分者間斯皮爾曼等級相關 18
表 9　評分者與入門基礎級八道試題評分偏誤頻率之統計表 20
表 10　標準化評分流程 22
表 11　試題難度分布 23
表 12　試題鑑別度分布 24
表 13　入門基礎級測驗整體模式適配度摘要表 26
表 14　自評問卷各題與測驗總分、通過等級之相關分析結果 28

圖目錄

圖 1　正式考試製卷流程 11

圖 2　評分流程 14

圖 3　入門基礎級測驗單因素驗證性因素分析 25

圖 4　入門基礎級測驗二因素驗證性因素分析 26

附件目錄

附件 1　入門基礎級口語測驗標準設定研究問卷調查結果 33

附件 2　各評分教師於不同試題的評分嚴格度分析結果 34

附件 3　華語文口語能力問卷-入門基礎級 36

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）[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（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。已在 2013 年於臺灣地區推出進階高階級正式考試，並於 2014 年推出入門基礎級正式考試。

本報告分為三部分，首先將針對入門基礎級口語測驗的內容、測驗實施與成績公布之標準化流程進行概述；其次，闡述入門基礎級正式考試之信、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。

[1] 流利精通級口語測驗已規畫於 2015 年舉辦預試，2016 年推出正式考試。

二、 簡介

2014 年度本測驗正式考試等級含入門基礎級（Band A）和進階高階級（Band B），依照測驗成績可區分為入門級（Level 1）、基礎級（Level 2）、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 A1（Breakthrough）、A2（Waystage）、B1（Threshold）與 B2（Vantage）。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於 2013 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）；以下針對入門基礎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，A1 等級學習者能針對與個人生活密切相關的信息，使用熟悉的日常用語與詞彙，進行簡單且制式化的回答；A2 等級學習者則能進一步針對日常生活中熟悉的事物及每日例行性事務，使用簡單的話語或句子進行描述，但短語或句子之間僅能使用簡單的連接詞或採用條列的方式連貫呈現；此外，也開始能簡短敘述一個故事。而在互動能力與溝通策略的部分，A1 等級學習者能在對方說話緩慢、清晰、簡單敘述的前提下，對於有即時需求的事物或熟悉的話題，例如，人物、地點及物品，使用簡單的短語及詞彙，詢問或回答簡單的問題，但溝通完全依賴速度緩慢地重覆、重述以及修正；A2 等級學習者則是能針對日常生活中，與工作、閒暇和例行事務相關的熟悉話題，處理非常簡短的社交往來，也能進行簡單直接的想法或資訊交換，例如，問路、指引方向及購買車票等基本信息，但對話無法持續進行。

本測驗研發人員（以下簡稱研發人員）綜合上述 CEFR 針對 A1、A2 等級所提出的口語能力描述，訂出入門基礎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
入門級	1. 能簡短地回答與個人生活密切相關的問題。例如：住在哪裡、認識什麼人、擁有的事物等。 2. 能使用熟悉的日常用語與詞彙簡單地描述人物、地點及物品。
基礎級	1. 能使用簡單的短語或句子敘述個人背景、日常生活中熟悉的事物及每日例行性事務。 2. 能簡單地描述短片的內容。

（二） 測驗題型及題數

研發人員據 CEFR 擬訂出入門級、基礎級的口語基本能力描述（如表 1）後，即循此方向設計測驗題型。

語言學習者在入門基礎級階段所發展的口語表達能力為描述性能力。學習者於入門級時，「能使用熟悉的日常用語與詞彙，進行簡短地敘述」，發展至基礎級時，學習者能進一步「使用簡單短語或句子，運用簡單連接詞或條列方式，進行連貫描述」；因此大致可將入門基礎級階段語言學習者的口語表達能力分為「單句層次描述能力」和「段落層次描述能力」二大面向；入門級學習者主要發展「單句層次描述能力」，至基礎級時，隨著段落層次組織及連貫能力的發展，學習者的描述能力已由單句層次延伸為段落層次，進一步展現出稍具穩定度的「段落層次描述能力」。

據此，在入門基礎級題型架構的設計上，針對「單句層次描述能力」和「段落層次描述能力」分別規劃了回答問題類和描述類兩大題型；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，著重於提供與個人生活密切相關的信息、簡單回答與工作、閒暇及日常例行生活有關的問題，並描述自身經驗以及對事物的喜好。

回答問題類的題型著重於受測者能否「以完整、簡短的句子回答問題」。描述類的題型則依據試題內容的素材，再細分為「經驗描述題」和「影片描述題」，前者著重於評量受測者能否「使用基本、簡單的短語、句子或常用的固定表達形式來描述所處環境中熟悉的日常事物或學習經驗」，後者則透過受測者是否能簡單地描述影片中所呈現的主題和重要內容，來評量受測者能否「使用基本、熟悉的短語、句子或固定的表達形式，具段落層次地簡單描述一個完整的事件」。

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。入門基礎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
入門基礎級	熱身題	2
	回答問題	4
	經驗描述	3
	影片描述	1

另外，在作答時間的制定上，本測驗參考了美國 AP 中文考試（The AP Chinese Language and Culture Exam）、中國漢語水平考試（Hanyu Shuiping Kaoshi，簡稱 HSK）、臺灣全民英檢（General English Proficiency Test，簡稱 GEPT）及法國法語鑑定文憑（DELF-DALF）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及內容完整度，最後制定出入門基礎級測驗的回答問題類題型，每一題的作答時間皆為 30 秒，描述類題型的每一題作答時間皆為 1 分鐘。

（三） 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」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，反映的就是受測者的單句層次描述性能力與段落層次描述性能力；以單句層次描述類型的語言任務為例，因單句層次的描述能力自入門級萌芽發展至基礎級已趨成熟，因此，通過入門級的語言使用者需「能使用非常簡單、基本、單獨的詞彙或短語，回答與個人密切相關且具體的問題」，通過基礎級的使用者則需能「使用簡單、基本的句子或常用的固定表達形式，簡短地回答與個人生活及例行事務相關的一般性問題」。至於，段落層次描述類型的

語言任務，因通過入門級的語言使用者尚不具備句子層次的組織及連貫能力，需至基礎級才具備有運用簡單連接詞，或條列方式連貫呈現話語內容的能力；故通過入門級的口語表現為「能使用熟悉的日常用語及詞彙，簡單地描述具體且熟悉的人物、地點及物品」，通過基礎級的語言使用者則需「能使用簡單、基本的句子或常用的固定表達形式，透過簡單的連接詞或條列方式，成串地連貫描述個人背景、日常生活中熟悉的事物、每日例行性事務及一個故事」。「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。

綜上所述，回答問題類題型考察的重點為考生單句層次的描述能力，描述類題型所考察的重點則為考生段落層次的描述能力；然而，因回答問題類題型的單句語料性質未涉及內容組織性及連貫性，詞彙和語法訊息有限，且句子之間的停頓、語詞重複性等因素也非單句層次所能考察的口語表達能力；因此，回答問題類題型的評分規準不需要也不適合像描述類題型一樣，將評分規準的內容細分為內容組織、語言運用和表達能力三個向度；這兩大類題型因而無法共用評分規準，而必須各自規劃對應的內容。

據此，研發人員針對回答問題類和描述類這兩大測驗題型各自規劃了一套評分規準，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）等方面，共同制定出入門基礎級口語測驗回答問題類和描述類題型評分原則，如表 3、表 4 所示，回答問題類評分原則適用於入門基礎級的「回答問題」題型，而描述題評分原則適用於「經驗描述」和「影片描述」這兩個題型。

本測驗採整體式評分，回答問題類題型與描述類題型的評分級距皆設定為 0 至 3 級分，評分者聆聽受測者的音檔內容後，再依據評分原則的內容，給予一整體等級分數。

表 3　入門基礎級回答問題評分原則

級分	評分原則
3	回答內容符合題目要求；語速適中，偶有停頓；詞語重複次數較少，幾乎都能被聽者理解；詞彙、語法掌握大致適當，偶有錯誤。
2	回答內容符合題目要求；語速緩慢，常有停頓，尚能被聽者理解；詞彙、語法簡單、掌握尚可，仍常有錯誤。
1	回答內容符合題目要求；語速過慢，停頓次數多且時間長，表達費力，較難被聽者理解；詞彙、語法簡單、掌握稍差，句子零散、破碎。 ※回答內容不完整或沒有直接回答。
0	離題；考生靜默，沒回答或等同未回答。

表 4　入門基礎級描述題評分原則

級分	內容組織	表達能力	語言運用
3	回答內容符合題目要求；內容尚稱充足，話語尚有組織。	語速適中，偶有停頓；詞語重複次數較少；語音尚稱清楚，偶有錯誤，幾乎都能被聽者理解。	能掌握基本詞彙和簡單的語法結構，使用大致適當，偶有錯誤。
2	回答內容符合題目要求；內容稍嫌不足，話題稍嫌無法擴展，組織較差。	語速緩慢，常有停頓；詞語重複次數多；部分語音不正確，尚能被聽者理解。	詞彙仍有限，使用尚稱適當，語法結構簡單、掌握尚可，仍常有錯誤。
1	回答內容符合題目要求；內容不足，組織差。 ※回答內容不完整或沒有直接回答。	語速過慢，停頓次數多且時間長，詞語重複次數過多，表達費力；語音多不正確，較難被聽者理解。	詞彙有限，使用多不適當，語法結構簡單、掌握稍差，句子零散、破碎。
0	離題；考生靜默，沒回答或等同未回答。		

（四） 通過門檻

本測驗透過標準設定（standard setting）程序，設定出入門級與基礎級之通過門檻。由於入門基礎級口語測驗給分方式為 0 至 3 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。

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. 說明回答問題類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. 請成員依據提供的入門級、基礎級最低能力描述，分別與回答問題類題型的評分原則進行配對，決定入門級和基礎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. 請成員聽完回答問題類題型的 10 個應試者音檔後，依據入門級、基礎級最低能力描述，分別判斷每個音檔的 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 及 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）每個音檔被判定為 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音檔進行第二回合門檻設定判斷，判斷方式同步驟 4 及 5 所示。
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 所示。
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回答問題類題型之通過門檻。

設定入門基礎級描述題之通過門檻時，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（見附件 1）共有 11 題四點量表，平均分數都在 3.36 以上。上述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，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（Spearman rank order coefficient）作為依據。標準差部分，從表 5 可知，入門級通過門檻部分，回答問題類題型的標準差在第一回合最大，然後逐漸降低；描述題可能經過上午回答問題的討論後，專家們判斷原則漸趨一致，並在第三回合 11 位專家的判斷達到完全一致，標準差為 0。基礎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
入門級	回答問題	0.522	0.302	0.302
	描述題	0.405	0.405	0.000
基礎級	回答問題	0.522	0.405	0.405
	描述題	0.467	0.405	0.000

由於口語測驗給分採級分制，由評分教師評閱後給予等級，屬於等級變項，因此採用斯皮爾曼等級相關分析回答問題與描述題各 10 個音檔 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。

回答問題與描述題各 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 A1、A1、A2 分別編碼為 0、1、2，與音檔實際級分求相關的結果，11 位專家回答問題題三個回合的相關係數依序為.838 至.937（p<.01）、.881 至.937（p<.01）、.881 至.937（p<.01），描述題三個回合的相關係數依序為.812 至.937（p<.01）、.896 至.937（p<.01）、.896 至.937（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皆具有高度的正相關存在，判斷結果與實際得

分頗為一致。

入門基礎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，即驗證了入門基礎級口語測驗能有效將華語學習者的口語表現區分為 CEFR 的 A1 和 A2 兩等級。

入門基礎級口語測驗的計分題目共有八題，各題均採 0 至 3 級分的評分級距，考生測驗總分為八題計分題的成績加總，滿分為 24 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 8 至 15 分者，可取得入門級（Level 1）證書，總分介於 16 至 24 分者，可取得基礎級（Level 2）證書。

表 6　入門基礎級口語測驗通過門檻分數

測驗等級	證書等級	分數範圍
入門基礎級	基礎級	16-24
	入門級	8-15

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。

2014 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：

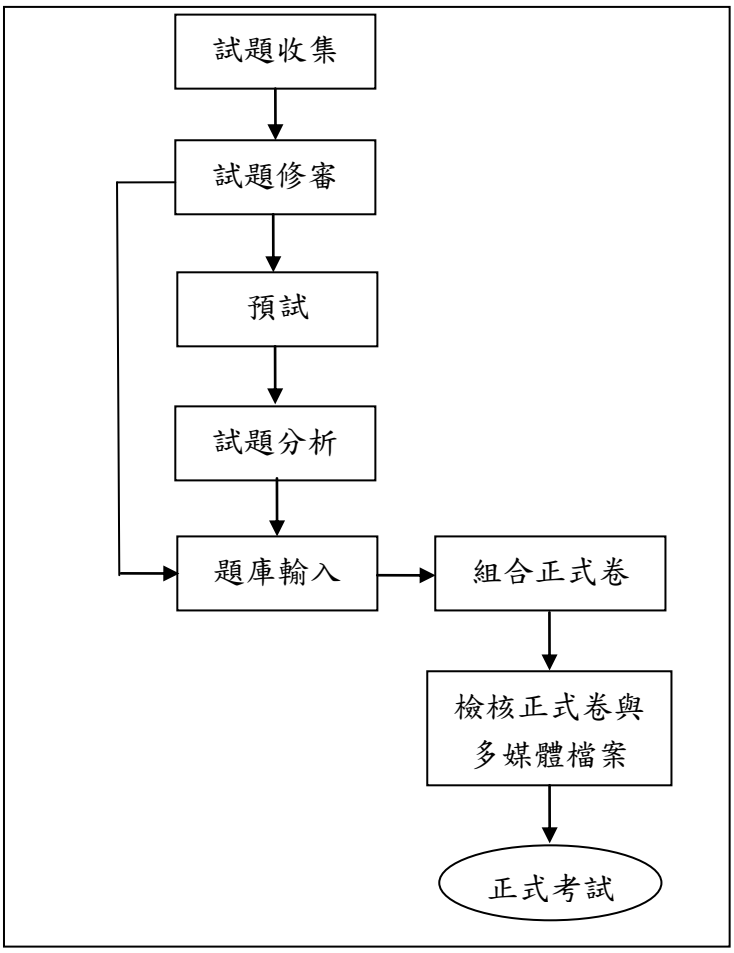

圖 1　正式考試製卷流程

1.　試題收集

2014 年度本測驗命題者共計九位（含入門基礎級、進階高階級），每位命題者每期繳交一套試題，每期命題時長約兩個月。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者命題指導文件及《華語八千詞詞表》[2]。根據年度統計，本年度的命題回收數量為：入門基礎級 148 題、進階高階級 104 題，共計 252 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。

[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

(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，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。本會於 2014 年 3 月 29 日舉辦全國性入門基礎級口語測驗預試。到考人數為 75 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.14 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（入門基礎級為 0-3 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（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，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 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。

2014 年入門基礎級預試共計八道試題，試題分析結果顯示，所有試題之 Infit MNSQ 與 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。此外，考量口語測驗題數較少，若所有試題皆需經由預試階段，較容易有外洩之虞，故口語測驗題庫之試題來源可分為兩種：經步驟 2 修審程序的完成之試題，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 64 題，分別為入門基礎級 35 題、進階高階級 29 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 2014 年 11 月 2 日舉辦入門基礎級、進階高階級口語測驗正式考試各一場；入門基礎級題型與題數如表 2 所示，進階高階級題型與題數詳見 2013 年口語測驗技術報告。

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：

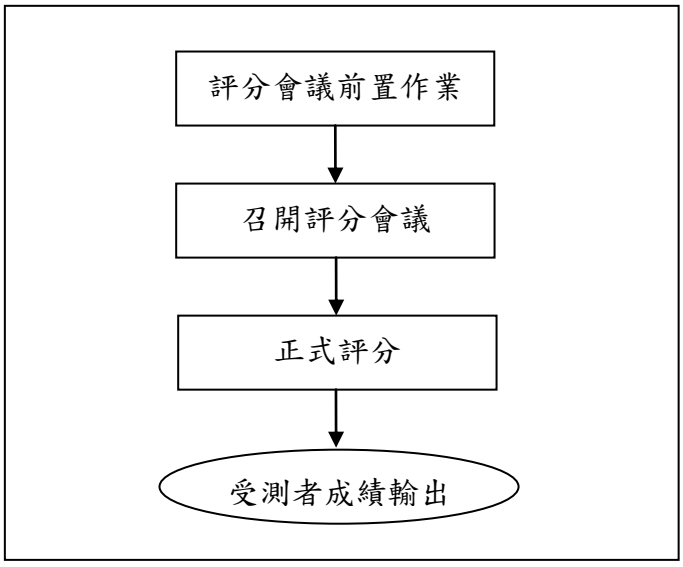

圖 2 評分流程

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。2014 年度召開的第一次口語測驗正式考試評分會議，共有十六位評分者參加：入門基礎級七位、進階高階級九位。

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二至三位評分者進行評分，評分方式採整體式評分（Holistic Scoring），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，即交由第三位評分者評定成績，其餘以此類推。成績處理上，則以所有評分者給分之眾數做為最終成績。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有兩名以上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。

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 2014 年 11 月 2 日所舉辦的全國性入門基礎級口語測驗正式考試之信、效度來說明本測驗之效能。

此次入門基礎級口語測驗正式考試的實際到考人數為 109 人。由七位評分者和一位研發人員參與評分，七位評分者分別評閱 49 至 50 名受測者之答題音檔，研發人員則評閱了全數 109 名受測者之答題音檔。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 2014 年入門基礎級口語測驗正式考試之信度。其中，將以「評

分者嚴格度變異」來評估評分者內一致性，並以「斯皮爾曼等級相關」來評估評分者間一致性。

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，八位評分者（含一位研發人員）中，有七位評分者的嚴格度約略落在±0.3 logit 以內，符合標準；評分者 A10 的嚴格度為 0.425，較為嚴格。

在評分者自身給分穩定性上，由嚴格度標準誤可發現，各評分者的標準誤介於 0.057 至 0.093 之間，顯示出八位評分者給分均具有自身的穩定性，其中 A16 為本會研發人員，評分規準掌握度較佳，故標準誤較其他七位評分者小。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，所有評分者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	Infit MNSQ	Outfit MNSQ
A10	49	0.94	0.425	0.093	1.07	1.04
A04	50	1.25	0.169	0.087	0.95	0.87
A08	49	1.05	0.083	0.090	0.83	0.77
A11	50	1.31	0.026	0.084	0.92	0.91
A02	50	1.35	-0.100	0.084	0.88	0.86
A17	49	1.11	-0.124	0.088	0.79	0.80
A05	50	1.36	-0.176	0.086	1.25	1.28
A16	109	1.43	-0.304	0.057	0.87	0.84

註：觀察的平均值表示評分者平均給分成績，因非所有評分者皆評閱相同考生，故不盡然平均給分較低者最為嚴格，反之亦然。

2. 評分者間信度

針對三組共八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。細格內數值為每一組兩兩評分者間相關係數之平均值，第一組評分者有三人，編號為 A02、A11、A16；第二組評

分者有三人，編號為 A04、A05、A16；第三組評分者共計四人，編號為 A08、A10、A16、A17。其中，A16 為本會研發人員，評閱所有受測者音檔。

第一組評分者各題平均值介於.770 至.916 之間，第三組評分者各題平均值則介於.779 至.918 之間，評分者間信度皆為大致良好；第二組評分者八道試題中有七道試題平均值達.8 以上，僅一題相關係數在.7 以下，但仍達到.01 之顯著水準。

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	第七題	第八題
第一組	.852	.916	.770	.784	.875	.910	.869	.853
第二組	.870	.826	.627	.847	.816	.949	.889	.839
第三組	.886	.802	.779	.788	.836	.901	.918	.807

由上述可知，2014 年入門基礎級口語測驗正式考試之評分者信度顯示，87.5%評分者嚴格度較為接近，而 A10 給分略為嚴格。在評分者內一致性方面，所有評分者皆符合適配度標準，且自身變異未過大。在評分者間一致性方面，則為多數評分者間信度大致良好，斯皮爾曼等級相關具有中度至高度的一致性。

為了確保評分者的評分品質，針對評分結果較差，如偏嚴格、偏寬鬆或一致性表現較不理想之評分者，本會透過第二次評分會議進行評分再訓練（P.22，表 10）。此外，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，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者列入觀察名單，若後續評分狀況仍未改善，即不續聘。

3. 研究子題－評分者評分偏誤研究

透過前述評分者內信度與評分者間信度的分析，發現所有評分者於評分者內一致性方面皆符合適配度標準；然據前人研究顯示（Eckes, 2008, 2012; Schaefer, 2008; Johnson & Lim, 2009），在表現型測驗中，即便是富有經驗，經嚴謹訓練過，且評分者間信度與評分者內信度表現皆良好的評分者，於其自身一致性內仍可能存在著評分者偏誤（rater bias），且其偏誤情形值得探討；此外，據研發人員觀

察，就過去幾年的評分結果、評分會議的討論情況來看，少數評分者似乎會受試題所指向的口語表達能力難易不同的影響，而在對不同試題評分時，出現給分略微偏寬鬆或偏嚴格的現象。因而，研發人員於本節中將進一步探討評分者內一致性的變異情形，包含評分者內是否存在著顯著性的評分者偏誤，以及可能造成評分偏誤的因素。

因過去評分者偏誤的相關研究，使用多面向 Rasch 測量模式可探討評分者與評分向度（category）、主題（topic）等的交互作用，而本研究欲探討評分者在面對不同試題時，是否存在不同的評分嚴格度，適用於此一分析方式，故採用可分析多面向 Rasch 測量模式之 Facets 3.71.3 版的多面向部分給分模式，透過分析八位評分者（含一位研發人員）於 2014 年 11 月入門基礎級口語測驗八道試題的評分資料，探討評分者與試題間是否存在交互作用，檢視各評分者內部評分的一致性是否受試題類別因素影響，而在不同試題間存在評分者偏誤。

今參考 McNamara（1996）研究作法，對試題與評分者交互作用分析結果的偏誤量（bias size）進行 t 考驗，以 t 值範圍介於-2.0 至 2.0 作為判斷標準，若超出此標準則表示存在顯著的評分偏誤，將八位評分者與入門基礎級八道試題評分結果間的偏誤頻率統計如表 9（因篇幅有限，各評分者於不同試題的評分嚴格度分析結果詳見附件 2）；表格左上方「S」表示「評分者於該試題給分比八道試題的整體性給分低，也就是內部給分偏嚴格」，同理，「L」則表示內部給分偏寬鬆；內部細格則以「數值/數值」的方式呈現給分偏嚴格或偏寬鬆的情形。舉例來說，編號 A04 評分者於第一題細格內所對應的數值為「1/0」，即表示 A04 第一題的給分標準相較於自身給分的整體標準而言，為一偏嚴格的情形；若評分者於該試題給分未出現偏嚴格或偏寬鬆的情形，對應的細格內將不會出現任何數值資料。

由表 9 的分析結果可知，八位評分者中，有五位評分者出現自身評分一致性會隨著試題的不同而有時出現偏寬鬆或偏嚴格的情況，其中，評分者評分偏誤頻率較高者為 A11 與 A10 兩位評分者，A10 評分者八道試題中有六道試題隨著個別試題的類別或難度不同，而出現內部一致性偏嚴格或偏寬鬆的情形，A11 評分者則出現八道試題中有三道試題內部一致性偏嚴格或偏寬鬆的情形；A04、A05 和 A17 三位評分者的偏誤情形次之，各僅有 1 至 2 題出現偏嚴格或偏寬鬆的情形。再由試題類別對評分者評分偏誤的影響來看，在回答問題類題型中給分偏寬鬆的情形有 3 例，偏嚴格的情形有 2 例，共計 5 例；在描述類題型中給分偏寬鬆

的情形有 3 例，偏嚴格的情形則有 6 例，共計 9 例，顯示出，評分者偏誤的情形似乎在描述類題型中較易發生。

另外，由各評分者在不同試題嚴格度的標準誤來看，各評分者對八道試題的標準誤介於 0.153 至 0.390 之間，顯示出八位評分者於進行每一道試題給分時，內部均具有一定穩定性，其中 A16 評分者可能由於為本會研發人員，較能掌握評分規準，八道試題的標準誤皆較小，介於 0.153 至 0.177 之間，而 A10 的標準誤較其他評分者大，介於 0.230 至 0.390 之間，其他六位評分者自身對八道試題評分的穩定性大致相當，皆介於 0.210 至 0.279 之間。再由 Infit MNSQ 介於 0.5 到 1.5 的標準，評估評分者於各道試題給分一致性是否如模式所預期，結果顯示，多數評分者之評分者內一致性均佳，自身評分穩定性良好。

表 9　評分者與入門基礎級八道試題評分偏誤頻率之統計表

題號（S/L） 評分者	回答問題類題型				描述類題型				
	第一題	第二題	第三題	第四題	第五題	第六題	第七題	第八題	共計
A02									0/0
A04	1/0				0/1				1/1
A05			0/1	1/0					1/1
A08									0/0
A10			0/1	0/1	1/0	1/0	1/0	1/0	4/2
A11					0/1		1/0	1/0	2/1
A16									0/0
A17								0/1	0/1
共計	1/0	0/0	0/2	1/1	1/2	1/0	2/0	2/1	8/6

由上述可知，2014 年入門基礎級口語測驗正式考試之評分者偏誤研究顯示，八位評分者中，評分偏誤頻率較高的評分者有兩位，分別為 A10 與 A11，研發人員推測，A10 評分者因近年教學經驗以高級班為主，故對入門基礎級口語能力的掌握較不穩定，導致該評分者整體內部一致性雖適配，但自身變異性過大，八道試題中有六道試題給分標準不一致；另一位評分者 A11 自身給分一致性的不穩定情況，則集中在描述題類型的試題，試題難度較高的描述類試題給分偏嚴，試題難度較低的描述類試題給分則略為寬鬆，回答問題類試題的自身評分一致性則表現穩定；研發人員據 A11 評分者提供的評分依據推測，該評分者於

評分過程，在考量受測者內容組織向度中關於段落層次、任務完成度的表現時，會不自覺依試題難度調整給分。

未來，將針對評分者個別評分偏誤的情形，進一步與各評分者溝通，並做為下次評分會議對該評分者訓練的重點。

（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、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 2014 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 10。

表 10　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，入門基礎級測驗之八位評分者嚴格度雖然有所不同，如評分教師 A10 嚴格度略為嚴格；然而，八位評分者中有七位評分者嚴格度差異約在±0.3 logit 以內，顯示絕大多數的評分者嚴格度符合標準，且所有評分者皆具有自身評分一致性（詳見表 7），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。探究評分者 A10 評分結果略為嚴格的原因：A10 評分者雖有一定次數的評分經驗，也參與了第一次評分會議，但在運用評分原則進行正式評分時，於「內容組織」此一向度的判定偏嚴，因而造成嚴格度差異較大的現象。此位評分者皆已於第二次評分會議中調整其嚴格度，未來將持續追蹤此評分者的評分嚴格度，並即時就其評分品質與評分者溝通討論，以期縮小評分者評分嚴格度的差異。

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，因第三題與第七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 11 所示，第二題與第六題偏難，第五題與第一題稍難，難度參數分別為 2.580、2.550、1.292 以及 1.071；第三題稍微容易，難度為-0.726；八道試題估計標準誤差異不大（介於 0.01 之間）。本測驗又採用 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（亦即超出範圍為題目不符合單向度試題反應理論模式），結果如表 11，試題與模式的適配情形皆良好，八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，意即入門基礎級口語測驗正式考試具有一定程度的建構效度。

表 11 試題難度分布

試題編號	難度	標準誤	Infit MNSQ	Outfit MNSQ
第二題	2.580	.075	0.85	0.81
第六題	2.550	.072	0.85	0.61
第五題	1.292	.070	0.82	0.89
第一題	1.071	.068	1.28	1.24
第七題	0.537[A]	.069	0.88	0.95
第八題	0.059	.079	0.63	0.61
第四題	0.037	.074	1.13	1.08
第三題	-0.726[A]	.077	0.88	1.06

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 12 所示，八道試題相關係數介於.713 至.884 之間，一般來說，鑑別度.40 以上表示非常優良（郭生

玉，2000），入門基礎級八道試題的鑑別度均在.40 以上，表示鑑別度良好。

表 12　試題鑑別度分布

試題編號	積差相關係數
第一題	.749
第二題	.713
第三題	.771
第四題	.757
第五題	.858
第六題	.806
第七題	.725
第八題	.884

(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。由於入門基礎級測驗包含回答問題類題型與描述類題型這兩大類題型，欲分別測量「單句層次描述能力」與「段落層次描述能力」，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，入門基礎級測驗雖包含兩類題型，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，入門基礎級測驗的測量變數為八道試題，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109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，入門基礎級單因素模式分析結果（如圖 4），因素負荷量介於 0.75 至 0.93 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。而二因素模式分析結果（如圖 5），因素負荷量介於 0.76 至 0.94 之間，因素負荷量標準誤同樣也達到顯著水準（p<.05）。此外，描述能力與說明能力二潛在因素之間的相關係數為 0.97；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入門基礎級口語測驗。

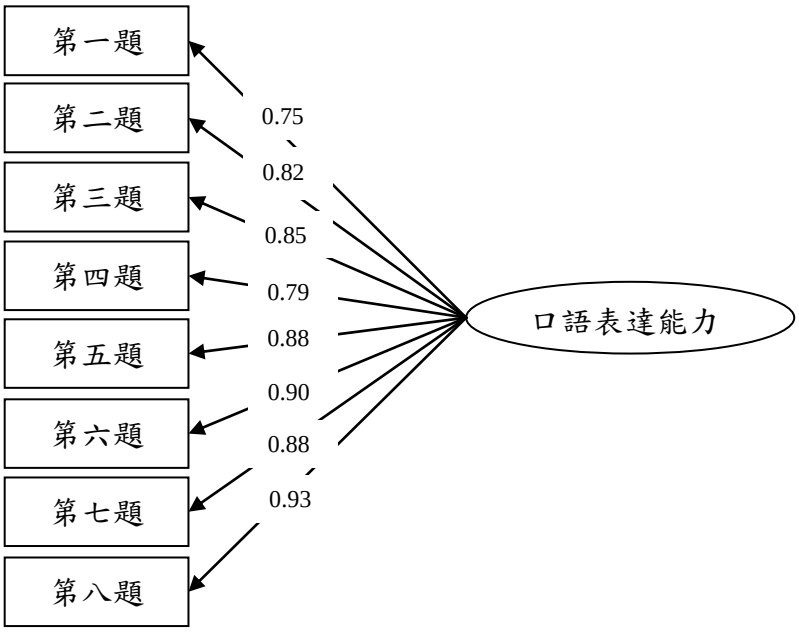

圖 3　入門基礎級測驗單因素驗證性因素分析

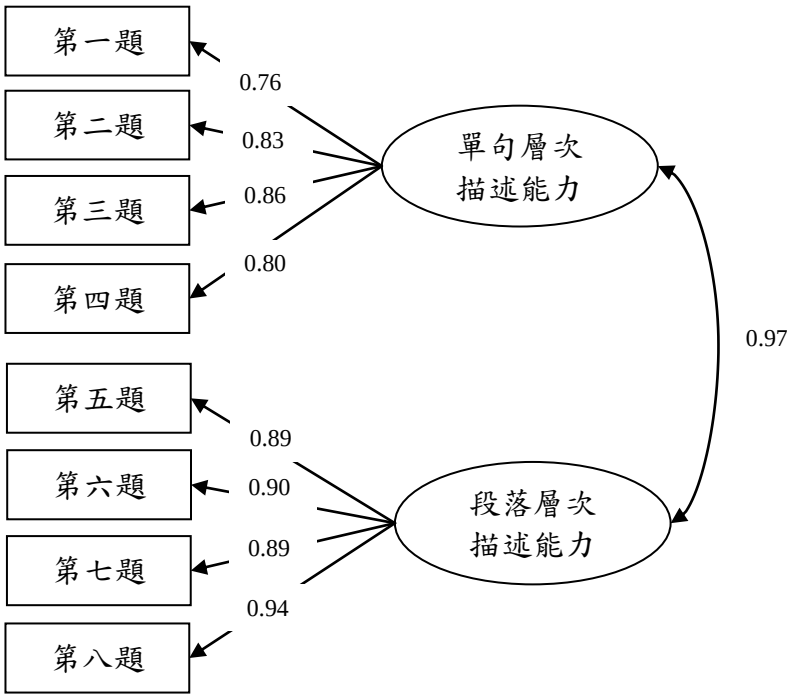

圖 4　入門基礎級測驗二因素驗證性因素分析

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，入門基礎級單因素與二因素模式的 χ^2 值分別為 17.47（p=0.62）及 16.40（p=0.63），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。檢視卡方自由度比結果，單因素與二因素模型均小於 3，亦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 數值均小於.08，表示兩種模式皆符合絕對適配度指標。增值適配度評估部分，CFI 和 NNFI 數值皆大於.90，顯示兩種模式均符合增值適配度指標。

表 13　入門基礎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	17.47	0.87	0.00	1.00	1.00
二因素模式	16.40	0.86	0.00	1.00	1.00

綜合上述結果可知，入門基礎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「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（如附件 3 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 14 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。14 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、測驗通過等級（不通過標記為 0；通過入門級標記為 1；通過基礎級標記為 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.575（p<.01），受測者自評結果與測驗通過等級的等級相關係數為.534（p<.01），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中度正相關存在，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 14 所示。自評問卷中，所有問題的答題反應與測驗總分的相關係數，皆達到.01 的顯著水準，相關係數介於.293 至.553 之間，表示在這 14 題中回答可以做到頻率越高的考生，其測驗總分也越高。而與通過等級的相關，同樣全部達顯著水準（p<.01），數值介於.290 至.504 之間，表示在這 14 題回答可以做到頻率越高的考生，通過測驗等級也越高。

整體來看，相關係數較低的為 Q3、Q12 和 Q13，Q3 這一題可能因能否使用「簡單的句子」為自評重點，而回答內容是否達到「句子」層次非入門基礎級考生能具體掌握的規範，因此造成相關係數偏低；Q12 這一題的題目內容中的自評重點為「談談曾經上過的學校、最近學校的情況」，可能因題目要求不夠明確，考生無法釐清需自評的項目是「學校」這個具體物件的描述，或「學校生活」這類與個人生活密切相關的經驗描述，而後者才是入門基礎級較能掌握的領域，因

此造成相關係數偏低的結果；至於 Q13 這一題則可能因其自評重點為「解釋喜歡和不喜歡什麼」，該主題較為抽象，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 Q4、Q5、Q6 和 Q8。其中，Q4、Q5 和 Q6 這三題相較於其他題目而言，題目本身的敘述中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，且自評重點皆為個人密切相關的經驗，可能因此使得自評結果與成績間的相關係數較高。至於 Q8 這一題的自評重點為「能否進行簡單故事或事情的簡短描述」，可能因該項為入門基礎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 14　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
測驗總分	.474**	.407**	.303**	.475**	.484**	.493**	.444**
通過等級	.432**	.416**	.290**	.461**	.482**	.472**	.397**
有效樣本	109	109	109	109	109	109	109
	Q8	Q9	Q10	Q11	Q12	Q13	Q14
測驗總分	.553**	.436**	.447**	.361**	.293**	.303**	.452**
通過等級	.504**	.370**	.462**	.363**	.305**	.309**	.436**
有效樣本	109	109	109	109	109	109	109

註：Q1-Q14 表示自評問卷題號；*表示 $p<.05$；**表示 $p<.01$。

2014 年 11 月 2 日入門基礎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八位評分者，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，顯示測驗具有效標關聯效度。

五、 結論

本文為 2014 年華語文口語測驗技術報告，闡述內容主要著重兩個部分，第一部分為針對入門基礎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，並說明本測驗之研發、施測和評分之標準化流程。第二部分則主要針對於 2014 年度首次推出正式考試的入門基礎級口語測驗之信度與效度評估，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。針對評分嚴格度變異較大以及自身給分一致性不穩定的評分者，將進行評分再訓練並列入觀察名單，若日後評分結果未獲改善，即不予續聘或改聘為其他適合等級的評分者。此外，也進行評分者偏誤研究，以進一步提升評分結果之一致性。

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 2014 年度全國性入門基礎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，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，自評口語能力較高者，其測驗成績也較高。

綜上所述，本年度入門基礎級口語測驗，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

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。

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。

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 January 17, 2007, from http://www.coe.int/t/dg4/linguistic/Source/Framework_EN.pdf

Eckes, T.（2008）. *Rater types in writing performance assessments: A classification approach to rater variability*. Language Testing, 25, 155–185.

Eckes, T.（2012）. *Operational rater types in writing assessment: Linking rater cognition to rater behavior.* Language Assessment Quarterly, 9, 270–292.

Schaefer, E.（2008）.Rater bias patterns in an EFL writing assessment. *Language Testing, 25（4）*, 465-493.

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

Johnson, J. S., & Lim, G. S.（2009）. *The influence of rater language background on writing performance assessment*. Language Testing, 26（4）, 485-505.

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.

McNamara, T. F.（1996）. *Measuring second language performance*. New York：Longman.

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

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.

附件 1　入門基礎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。	3.82
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。	3.73
3.我了解最低能力者在標準設定方法上的涵義。	3.82
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。	3.91
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。	3.73
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。	3.91
7.我是根據 A1 最低能力描述判斷 A1 通過門檻分數。(程序 3)	3.70
8.我是根據 A2 最低能力描述判斷 A2 通過門檻分數。(程序 3)	3.64
9.我對於自己所設定的通過門檻分數（cut score）有信心。	3.36
10.我是根據 A1 最低能力描述判斷考生音檔 CEFR 等級。(程序 4)	3.73
11.我是根據 A2 最低能力描述判斷考生音檔 CEFR 等級。(程序 4)	3.73

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　各評分教師於不同試題的評分嚴格度分析結果

觀察分數	期望分數	評閱筆數	觀察分數與期望分數差值之平均	偏誤模式			加權適配度統計值	評分者			試題	
				偏誤量	標準誤	t 值	均方	序號	編號	嚴格度	編號	難度
83	69.24	49	0.28	0.975	0.258	3.78	0.8	19	A10	0.425	3	-0.726
72	59.13	50	0.26	0.725	0.238	3.05	0.8	38	A04	0.169	5	1.261
78	65.86	49	0.25	0.713	0.244	2.92	1.8	27	A10	0.425	4	0.010
79	70.61	49	0.17	0.591	0.265	2.23	0.5	58	A17	-0.120	8	0.069
97	89.01	50	0.16	0.532	0.260	2.05	1.7	23	A05	-0.180	3	-0.726
72	62.42	50	0.19	0.501	0.229	2.19	0.9	36	A11	0.026	5	1.261
41	33.63	50	0.15	0.405	0.228	1.77	0.7	46	A04	0.169	6	2.487
93	86.92	50	0.12	0.384	0.251	1.53	0.8	20	A11	0.026	3	-0.726
21	17.05	49	0.08	0.376	0.295	1.28	0.9	11	A10	0.425	2	2.533
45	37.34	50	0.15	0.370	0.215	1.72	0.7	45	A02	-0.100	6	2.487
77	71.72	49	0.11	0.312	0.244	1.28	1.0	25	A08	0.083	4	0.010
76	70.20	50	0.12	0.298	0.228	1.31	1.7	5	A02	-0.100	1	1.027
85	81.36	50	0.07	0.254	0.264	0.96	0.7	63	A05	-0.180	8	0.069
80	76.44	50	0.07	0.250	0.265	0.94	1.0	62	A04	0.169	8	0.069
210	201.68	109	0.08	0.246	0.172	1.43	0.8	24	A16	-0.300	3	-0.726
49	44.51	49	0.09	0.240	0.230	1.04	0.8	3	A10	0.425	1	1.027
80	76.75	49	0.07	0.225	0.261	0.86	0.6	18	A17	-0.120	3	-0.726
59	54.93	49	0.08	0.211	0.227	0.93	1.0	2	A17	-0.120	1	1.027
192	185.57	109	0.06	0.200	0.177	1.13	0.6	64	A16	-0.300	8	0.069
88	85.82	50	0.04	0.136	0.251	0.54	1.1	28	A11	0.026	4	0.010
70	67.71	50	0.05	0.116	0.225	0.52	1.6	4	A11	0.026	1	1.027
42	39.87	50	0.04	0.111	0.227	0.49	1.1	47	A05	-0.180	6	2.487
202	199.44	109	0.02	0.074	0.170	0.43	1.2	32	A16	-0.300	4	0.010
32	31.23	50	0.02	0.051	0.255	0.20	0.8	14	A04	0.169	2	2.533
87	84.92	109	0.02	0.050	0.154	0.32	0.8	16	A16	-0.300	2	2.533
66	65.25	50	0.01	0.042	0.237	0.18	1.2	39	A05	-0.180	5	1.261
35	34.27	50	0.01	0.040	0.233	0.17	0.6	13	A02	-0.100	2	2.533
68	67.70	49	0.01	0.021	0.267	0.08	0.5	57	A08	0.083	8	0.069
74	73.83	49	0.00	0.012	0.269	0.04	1.0	17	A08	0.083	3	-0.726
65	64.84	50	0.00	0.008	0.229	0.04	0.8	37	A02	-0.100	5	1.261
22	21.90	49	0.00	0.008	0.279	0.03	0.6	41	A08	0.083	6	2.487
70	70.18	50	0.00	-0.010	0.235	-0.04	1.6	7	A05	-0.180	1	1.027
57	57.48	49	-0.01	-0.027	0.237	-0.11	1.1	49	A08	0.083	7	0.537
69	69.50	50	-0.01	-0.027	0.234	-0.12	0.6	54	A04	0.169	7	0.537
36	36.80	50	-0.02	-0.047	0.243	-0.19	1.0	15	A05	-0.180	2	2.533
34	34.90	50	-0.02	-0.048	0.231	-0.21	0.7	44	A11	0.026	6	2.487

觀察分數	期望分數	評閱筆數	觀察分數與期望分數差值之平均	偏誤模式			加權適配度統計值	評分者			試題	
				偏誤量	標準誤	t 值	均方	序號	編號	嚴格度	編號	難度
50	50.95	49	-0.02	-0.050	0.229	-0.22	0.8	1	A08	0.083	1	1.027
20	20.61	49	-0.01	-0.055	0.302	-0.18	0.9	9	A08	0.083	2	2.533
31	32.02	50	-0.02	-0.059	0.243	-0.24	0.6	12	A11	0.026	2	2.533
159	161.59	109	-0.02	-0.061	0.154	-0.40	1.3	8	A16	-0.300	1	1.027
60	61.24	49	-0.03	-0.067	0.234	-0.29	0.8	50	A17	-0.120	7	0.537
49	50.49	49	-0.03	-0.083	0.237	-0.35	0.7	34	A17	-0.120	5	1.261
171	174.98	109	-0.04	-0.094	0.153	-0.61	0.7	56	A16	-0.300	7	0.537
74	75.99	50	-0.04	-0.101	0.226	-0.45	1.0	53	A02	-0.100	7	0.537
87	91.71	109	-0.04	-0.109	0.153	-0.71	0.8	48	A16	-0.300	6	2.487
86	88.92	50	-0.06	-0.185	0.252	-0.73	0.5	21	A02	-0.100	3	-0.726
72	75.23	49	-0.07	-0.191	0.242	-0.79	0.8	26	A17	-0.120	4	0.010
142	150.10	109	-0.07	-0.195	0.155	-1.26	0.7	40	A16	-0.300	5	1.261
43	46.80	49	-0.08	-0.219	0.242	-0.90	0.7	33	A08	0.083	5	1.261
71	75.88	50	-0.10	-0.263	0.233	-1.13	1.3	55	A05	-0.180	7	0.537
83	87.84	50	-0.10	-0.301	0.248	-1.21	0.7	29	A02	-0.100	4	0.010
77	81.60	50	-0.09	-0.309	0.260	-1.19	0.6	61	A02	-0.100	8	0.069
76	81.39	50	-0.11	-0.348	0.253	-1.38	1.0	30	A04	0.169	4	0.010
20	24.69	49	-0.10	-0.362	0.292	-1.24	1.0	42	A17	-0.120	6	2.487
78	83.80	50	-0.12	-0.391	0.262	-1.49	0.9	22	A04	0.169	3	-0.726
18	23.07	49	-0.10	-0.452	0.316	-1.43	1.0	10	A17	-0.120	2	2.533
64	73.50	50	-0.19	-0.492	0.230	-2.14	0.9	52	A11	0.026	7	0.537
43	51.55	49	-0.17	-0.534	0.257	-2.08	1.1	51	A10	0.425	7	0.537
78	86.64	50	-0.17	-0.565	0.254	-2.22	1.1	31	A05	-0.180	4	0.010
71	79.71	50	-0.17	-0.589	0.262	-2.25	0.5	60	A11	0.026	8	0.069
31	40.92	49	-0.20	-0.633	0.263	-2.41	0.6	35	A10	0.425	5	1.261
54	62.96	49	-0.18	-0.667	0.277	-2.41	0.5	59	A10	0.425	8	0.069
51	63.88	50	-0.26	-0.700	0.234	-2.99	0.9	6	A04	0.169	1	1.027
11	17.92	49	-0.14	-0.823	0.390	-2.11	1.1	43	A10	0.425	6	2.487

附件 3　華語文口語能力問卷-入門基礎級

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	我能用簡單的詞彙說出人物和地點，但句子大多零散。 I can produce simple mainly isolated phrases about people and places.	1	2	3	4	5
2	我能描述我自己、我做的工作和住在什麼地方。 I can describe myself, what I do and where I live.	1	2	3	4	5
3	我能用簡單的句子談談我的家、我家人和我認識的人。 I can talk about my home, my family and people I know in simple sentences.	1	2	3	4	5
4	我能描述我平常做什麼事情。 I can describe what I do regularly.	1	2	3	4	5
5	我能描述我以前做過什麼事情。 I can describe what I did in the past.	1	2	3	4	5
6	我能描述我打算要做的事情和計畫。 I can describe plans, arrangements and alternatives.	1	2	3	4	5
7	我能大致說出我已經做過的事情。 I can tell the main points about something I have done.	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
8	我能簡短地描述一個簡單的故事或一件事情。 I can give short simple descriptions of events or tell a simple story.	1	2	3	4	5
9	我能簡單地描述事物而且做出比較。 I can give simple descriptions of things and make comparisons.	1	2	3	4	5
10	我能描述參加過的活動和我自己的經驗。（例如：上個週末、上個假日） I can describe past activities and personal experiences（e.g. the last weekend, my last holiday）.	1	2	3	4	5
11	我能描述我的教育背景、我現在或最近的工作。 I can describe my educational background, my present or most recent job.	1	2	3	4	5
12	我能談談我曾經上過的學校、我最近學校的情況或我最近的工作。 I can talk about which schools I've gone to, my present educational situation or my present job.	1	2	3	4	5
13	對於一件事情，我能解釋我喜歡什麼和不喜歡什麼。 I can explain what I like and don't like about something.	1	2	3	4	5
14	對於熟悉的話題，在能預先準備的情況下，我能做出簡短、基本的報告。 I can give a short, rehearsed, basic presentation on a familiar subject.	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—2014(3)
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：2016 年 11 月
定　　價：新台幣 100 元