兒童華語文能力測驗技術報告－2013(5)

聽力、閱讀測驗信效度

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

一、 前言

「兒童華語文能力測驗」(Children's Chinese Competency Certification) (以下簡稱兒童測驗)是一套專為七至十二歲母語非華語之兒童學習者所設計的標準化語言能力測驗，由「國家華語測驗推動工作委員會」(以下稱為本會)專責研發。兒童測驗旨在幫助兒童學習者了解自己的華語程度、激勵他們學習的熱忱，故不以任何特定教材為命題依據，測驗題材取自真實的生活情境，豐富有變化，測驗內容涵蓋所有與兒童經驗相關之主題，如日常生活與休閒活動、交通運輸、人物特徵、身體與健康、學校用語、時間與空間概念以及氣候與季節等。測驗實施方式為紙筆測驗，聽力測驗與閱讀測驗合併施測。

以下將先簡介 2013 年兒童測驗的能力描述、內容及各等級通過門檻，接著描述兒童測驗製卷與成績公布之標準化流程，再說明該年度正式考試信效度的分析結果，最後根據各項分析結果進行討論與建議。

二、 簡介

兒童測驗共分為三個等級，由易至難依序為萌芽級(Sprouting)、成長級(Seedling)與茁壯級(Blossoming)。測驗以「歐洲共同語文參考架構」(Common European Framework of Reference for Languages；簡稱 CEFR)之 A1 級(Breakthrough)和 A2 級(Waystage) (Council of Europe，2001)的語言能力描述及定義為架構，所評量的語言能力皆落在初級語言使用者(Basic User)的範圍之內。其中，成長級和茁壯級分別根據 CEFR 之 A1 級和 A2 級發展而來。與此同時，為了讓初學華語且年齡較小的兒童能夠順利銜接、熟悉測驗方式並建立自信，兒童測驗基於 A1 級能力之簡單語言任務描述再往下針對更為基礎的能力研發出萌芽級，並將萌芽級定為 Pre-A1 級(陳怡靜、趙家璧，2012)。此節將介紹各等級能力描述、測驗題型與題數以及通過門檻。

(一)能力描述

兒童測驗以 CEFR 初級語言使用者的能力描述為架構，發展出萌芽、成長、茁壯三級的能力指標，各等級通過測驗者具備的基本能力如表 1 所示。由於兒童測驗是針對兒童學習者所研發，因此，基本能力描述的制定限定在與兒童生活經驗相關的語言使用範疇之內。

為提供潛在受測者更精準的語言學習歷程及本測驗所測能力之相互參照，本測驗在 2013 年修訂各級測驗的能力描述，由先前偏向整體語言能力的描述方式，改由針對聽力與閱讀兩項語言能力進行分項描述，並對各等級間的能力差異做出區隔。

表 1 兒童華語文能力測驗能力描述

測驗等級	聽力理解能力	閱讀理解能力
萌芽級	在話語速度非常緩慢、發音清晰，且停頓或重複的前提下，能理解熟悉且基本的詞語，例如個人資訊、數字、顏色、天氣、喜好、問候語等。	能理解基本的詞語及簡單的句子。在視覺輔助的前提下，藉由掌握基本詞語，能大致理解簡單的書寫材料。
成長級	在話語速度慢、發音清晰時，能理解與日常生活有關的簡短談話，例如與家人的互動、學校生活、購物、穿著、飲食、交通等等。 在聆聽簡短、簡單、發音清晰的宣佈及說明時，能理解其中的要點，例如自我介紹、電話留言、課堂上的宣佈或天氣預報等。	在視覺輔助下，能理解簡單敘事短文的大意及找出重要訊息。能從日常生活的簡易書寫材料，例如菜單、車票、明信片及標示中，讀出基本訊息，例如：姓名、日期、時間、價錢、地點等。
茁壯級	能理解有關居家生活、興趣嗜好、旅遊、休閒活動、同儕間的互動等主題的談話。在聆聽發音清晰的說明或錄製片段時，能掌握主旨及重要資訊，例如：介紹、宣佈、廣播、新聞等。	能理解主題具體、與個人生活經驗相關的簡單敘事短文。能辨識日常生活中不同書寫材料的功能，且能從書寫材料，例如留言、便條、信件、學校公告、廣告或海報中，讀出重要的訊息，例如：活動時間、地點、辦法、適用對象、注意事項等。

(二)測驗題型與題數

由於本測驗的目標受測者為七到十二歲的兒童，此階段兒童的認知能力及語言能力尚在發展中，因此，研發題型時，需將兒童受測者的發展特性、語言使用範疇及其生活經驗等一併納入考量。首先，兒童專注時間較短(Mckay，2006)，在一次測驗中，需藉由多元的作答方式與題型的轉換，提升兒童受測者的注意力，以期能客觀地評估受測者真正的語言能力。其次，兒童生活經驗較為有限，故測驗內容所包含的主題、情境與任務需與兒童生活經驗相關，避免因為兒童沒有充足的背景知識而影響其作答。再者，兒童的測驗經驗較少，故須藉由兒童熟悉的測驗形式，以降低作答時因不熟悉測驗方式而產生焦慮，進而影響作答的可能性。此外，此階段受測者為初級語言使用者，整體能力發展相當初階，理解文

本的同時往往需要視覺輔助，因此無論是聽力測驗或是閱讀測驗都採用了多種圖文相輔的題型。

本會兒童測驗研發人員(以下簡稱研發人員)綜合考量兒童發展特性、語言使用範疇及生活經驗等因素，並依據三等級學習者不同的語言能力描述(見表 1)，共設計十一種題型來測量兒童華語文能力。其中，聽力測驗共有六種題型—聽力選圖(聽單句選圖、聽對話選圖)、看圖回答、聽力連連看(聽句子連連看、聽對話連連看)、完成對話、會話理解以及聽真實材料題；閱讀測驗共有五種題型—看圖辨別句義、圖文連連看、選詞填空、閱讀材料及短文理解。由於萌芽級與成長級受測者的語言發展能力尚處於初級語言者初階的階段，因此，在萌芽級與成長級聽力與閱讀測驗多採用圖文相輔的題型；茁壯級受測者的語言能力發展處於初級語言使用者較純熟的階段，故茁壯級聽力與閱讀測驗圖文相輔的題型較少，主要使用對話、短文或材料等需要較高理解能力方能作答的題型。表 2 為兒童測驗三等級的題型與題數分布，其中，萌芽級因考量該階段受測者初學漢字，在閱讀文字時產生的認知負荷量較大，故閱讀測驗所安排的題數少於聽力測驗，其餘兩等，在同一等級內的聽力、閱讀題數皆相同；三等級施測的總題數分別為萌芽級 40 題、成長級 50 題以及茁壯級 60 題，施測時間則分別為 40 分鐘、50 分鐘以及 60 分鐘。

表 2 測驗題型與題數分布

測驗等級	測驗內容	題型	題數	總題數
萌芽級	聽力理解	聽單句選圖 (選擇)	7	25
		看圖回答 (選擇)	6	
		完成對話 (選擇)	7	
		聽句子連連看	5	
	閱讀理解	看圖辨別句義 (是非)	5	15
		圖文連連看	4	
		選詞填空 (選擇)	6	
成長級	聽力理解	聽對話選圖 (選擇)	6	25
		看圖回答 (選擇)	6	
		完成對話 (選擇)	6	
		聽對話連連看	7	
	閱讀理解	圖文連連看	4	25
		選詞填空 (選擇)	6	
		閱讀材料 (選擇)	6	
		短文理解 (是非)	9	
茁壯級	聽力理解	聽對話選圖 (選擇)	7	30
		完成對話 (選擇)	7	
		對話理解 (選擇)	8	
		真實材料 (選擇)	8	
	閱讀理解	選詞填空 (選擇)	10	30
		閱讀材料 (選擇)	10	
		短文理解 (是非)	10	

(三)通過門檻

各等級測驗題目無論選擇題、是非題或連連看各題型，答對一題皆得一分；答錯不倒扣。兒童測驗屬於標準參照測驗(criterion-referenced test)，各等級通過門檻均設定為答對總題數達 60%以上者方可獲得證書。即萌芽級成績達 24 分以上者，可獲得萌芽級證書；成長級成績達 30 分以上者，可獲得成長級證書；而茁壯級成績達 36 分以上者，方可獲得茁壯級證書。

三、 測驗標準化流程

2013 年度兒童測驗標準化流程共包含兩部分(如圖 1 所示)，第一部分為進行正式考試前之正式考試製卷，共包含：試題的收集、修審、預試、分析、題庫輸入、組合正式卷、檢核正式卷與多媒體檔案等七個步驟；第二部分為考試後之成績公布，各部分詳細說明如下所述：

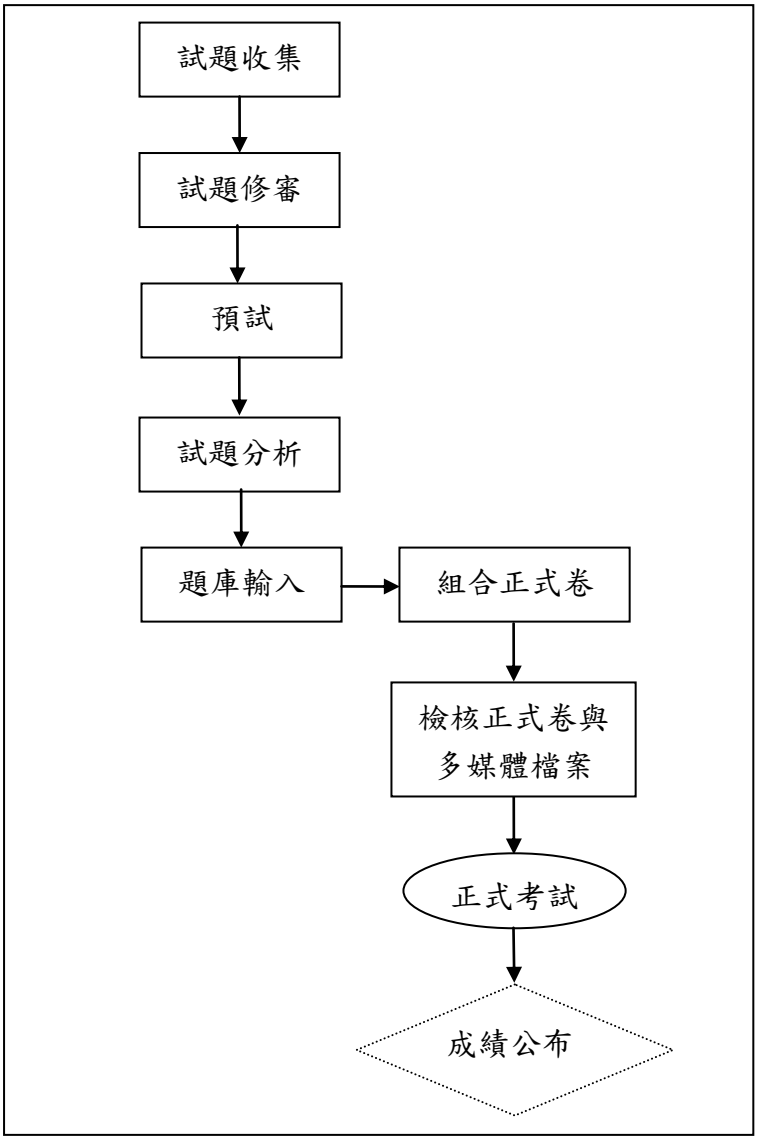

圖 1 測驗標準化流程

(一)試題收集

2013 年兒童測驗試題收集分為兩階段，先舉辦命題研習，並於研習後回收試題進行審查，聽力測驗與閱讀測驗共有十二位老師通過審查。另外再召開進階

命題座談會，由測驗研發人員彙整命題常見問題，回饋審題意見並加強與命題教師之溝通後，再次進行收題。本年度聽力測驗共回收 205 道題目，閱讀測驗則回收 235 道題目。

(二)試題修審

1. 會內初審

將試題收集階段完成修訂的試題，交由非兒童測驗研發人員共三人進行第一階段會內初審，並回收審題意見，審題時間為兩週。接著由研發人員根據審題意見修改試題，修改時間為兩週。

2. 專家學者外審

會內初審修改後的試題，在第二階段將交由華語文教學、測驗領域的專家學者及資深華語教師審查並提供意見，此階段審題時間約二至四週。研發人員再依據審題意見修改試題，修改時間為兩週。

3. 製作試題相關多媒體檔案

在進行預試前，審訂完成的試題需製作相關媒體檔案，包含製作測驗說明音檔及試題音檔，繪製並確認試題圖片及題本排版。在測驗說明與試題音檔製作方面，延請專業錄音員進行錄音，並由錄音室製作音檔，過程中研發人員負責確認錄音內容正確性與音檔品質。

此外，考量兒童測驗是為初級華語學習者而設計，為輔助受測者了解考試進行方式，測驗說明音檔的製作採華語搭配其他語言的形式，多國語言測驗說明版本包含中英版、中西版、中法版、中德版、中越版、中韓版以及中日版。

在此最終定稿階段，試題圖片如需更動修改，製作時間將延長一至兩週，待圖片定稿後，再交由美編人員排版，並由研發人員確認內容直至無誤，最後交由印刷廠商印刷預試題本。

(三)預試

為了提高試題難度估計之精準度及預試樣本收集之效率，自 2013 年起，兒童測驗預試改採三等合併之形式收集樣本。一份預試卷中包含為萌芽級、成長級及茁壯級受測者設計的試題，並將聽力測驗與閱讀測驗的題本各自獨立，考生可

選擇只參加一項測驗或兩項測驗皆參加。這項改變能使每道試題受測樣本的能力分布更為廣泛，降低試題難度參數[1]之估計誤差，提高估計精準度；亦可有效避免單一等級試題因預試樣本較少而無法完成收樣的情形。

本年度兒童測驗透過臺灣地區及海外地區收集預試樣本，其中聽力測驗預試受測者為 910 人次，閱讀測驗則為 886 人次。

(四)試題分析

將預試階段受測者之作答反應交由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。由於兒童測驗中各試題作答反應皆為非對即錯，屬於二元計分方式(dichotomous items)，故採取 IRT 中的 Rasch 模式(Rasch, 1960)進行資料分析，分析軟體為 Winsteps 3.68.2 版。Rasch 模式如公式(1)所示：

$$P(X_{ni}=1|\theta_n,\delta_i)=\frac{\exp(\theta_n-\delta_i)}{1+\exp(\theta_n-\delta_i)} \quad (1)$$

其中，θ_n 表示受測者 n 的能力；δ_i 表示試題 i 的難度(difficulty)；$P(X_{ni}=1|\theta_n,\delta_i)$ 表示受測者 n 在試題 i 答對的機率。由公式(1)可發現，Rasch 模式的假設為，受測者答對每一道試題機率受到受測者具備的能力及試題難度的影響，將受測者在各試題上的反應，透過公式(1)即可估計出受測者的能力以及試題的難易度，並且讓施測不同試題的受測者其能力可以互相比較。因為，在 IRT 中，受測者的能力參數與試題參數是同時納入在一個模式裡，因此，估計受測者能力時，已經考量了試題參數的影響，估計試題參數時，也考量了受測者能力的影響。

評估預試試題品質的依據為 Winsteps 3.68.2 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)，及其標準化 Z 值(z-standardized)，分別簡稱為 Infit MNSQ 及 Infit ZSTD。評估標準為試題之 Infit MNSQ 數值介於 0.7 至 1.3 者以及 Infit ZSTD 介於-3.0 至 3.0 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。同時，尚提供試題鑑別度指標[2]、難度 P 值[3]及選項分析等資料作為輔助，以更全面的觀點評估試題品質。

[1] 難度參數指的是由 Winsteps 軟體所估計出的試題難度參數估計值。
[2] 試題鑑別度指標為 D 值及點二系列相關係數。
[3] 難度 P 值為高低分組受測者答對率之平均數。

(五)題庫輸入

預試後的試題在經由試題分析後，由研發人員參考各評估指標、試題相關資料以及內容等進行討論，以決定試題是否保留。扣除共同試題(或稱定錨試題)後，本年度兒童測驗題庫新進試題題數為聽力測驗 68 題、閱讀測驗 56 題。

(六)組合正式卷

兒童測驗正式卷皆由經過預試並根據受測者作答反應進行試題分析之試題組成，各等級測驗組卷之平均難度設定以中等偏易為原則(即難度參數小於 0.5 logit 之試題居多)，並依照固定比例分配各難度區間之題數，自題庫中挑選難度參數符合試卷難度區間要求的試題，組成各測驗等級正式卷之主要試題難度區間範圍皆介於-3.00 至 1.00 之間。而三等級測驗試題各自的試題難度量尺分布大約都介於-3.00 至 3.00 之間。

正式卷初稿組卷完成後，由研發人員檢核各卷試題之主題分布是否過度集中、是否出現重複的考點、是否符合雙向細目表(參見頁 14-15，表 9 及表 10)之規劃等，再根據檢核建議更換試題，直至整份試卷皆無上述問題後始定稿。2013 年度各測驗等級正式考試卷之難度分布如表 3 至表 5 所示。

表 3 萌芽級正式考試卷難度分布

難度	聽力	閱讀
<-3.00	0	0
-3.00~-2.50	0	0
-2.49~-2.00	0	0
-1.99~-1.50	1	0
-1.49~-1.00	4	1
-0.99~-0.50	8	4
-0.49~0	3	4
0.01~0.50	5	5
0.51~1.00	2	0
>1.00	2	1
總計	25	15

表 4 成長級正式考試卷難度分布

難度	聽力	閱讀
<-3.00	1	0
-3.00~-2.50	0	0
-2.49~-2.00	0	0
-1.99~-1.50	2	0
-1.49~-1.00	5	7
-0.99~-0.50	3	4
-0.49~0	1	8
0.01~0.50	6	4
0.51~1.00	4	1
>1.00	3	1
總計	25	25

表 5 茁壯級正式考試卷難度分布

難度	聽力	閱讀
<-3.00	0	0
-3.00~-2.50	0	0
-2.49~-2.00	1	0
-1.99~-1.50	0	1
-1.49~-1.00	1	0
-0.99~-0.50	9	5
-0.49~0	6	10
0.01~0.50	8	9
0.51~1.00	4	3
>1.00	1	2
總計	30	30

(七)檢核正式卷與多媒體檔案

由於兒童測驗採紙筆作答的施測形式，因此，正式卷組卷完成後，必須印刷題本。交付印刷前，研發人員針對題本排版內容進行校對工作，首先為正體版題本的校對，包含確認錄音稿與題本的內容是否一致、檢查圖片與文字的搭配是否正確、漢語拼音與注音符號的標記是否正確、題號與頁數的編碼是否正確等。確認無誤後，再校對簡體版題本，內容及程序同正體版題本。此階段需反覆校對正、簡兩個版本與錄音稿件間的一致性，直至正確無誤。確認題本排版後，即根據錄音稿件完成錄音工作，並與排版完成之正、簡題本互相校對，確認錄音品質與內容正確性，確認後試卷檢核即完成，並交付印刷，作為正式考試之用。

兒童測驗成績公布流程如下：首先，檢核受測者作答資料並確認無誤後，接著檢核成績報表，確認無誤後即為受測者考試成績，最後進行成績單及證書之印製與寄發工作。

四、 測驗評估

一份測驗是否能夠發揮效用，並能確切地測量受測者的目標潛在能力，通常可透過該測驗的信度與效度分析來進行整體性的評估。因此，本節將討論 2013 年度兒童測驗之信效度來總結兒童測驗之效能。2013 年度兒童測驗採聽力和閱讀測驗合併施測的形式，雖然聽力理解能力與閱讀理解能力皆屬於語言能力中的接收能力(receptive skills)，兩者間具有一定程度的相關，但仍各有其獨特之處，故以下分析將分別描述聽力測驗與閱讀測驗結果。2013 年度正式考試三個等級分別都使用了兩卷次，各等到考總人數分別為萌芽級 1501 位、成長級 835 位，以及茁壯級 411 位。其中，萌芽級有 1197 人次，成長級 723 人次，與茁壯級 332 人次使用同一卷次，占本年度報考人數約八成，因此接下來將以此卷獲得之考生資料及作答反應為代表，分析測驗信效度。

(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，無論在什麼時間、什麼地點，由何人進行施測、計分，都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，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。

兒童測驗為一次性測驗，亦即測驗僅施測一次，所有受測者於同一時間接受同一份測驗，因此，可藉由內部一致性信度來表示測驗信度。當每道試題測量結果的相關性高時，則顯示測驗試題皆測量到相同的潛在特質，即內部一致性高；反之，當測量結果的相關性低，則表示測驗試題測量到其他潛在特質，即內部一致性低。常用於表示內部一致性的指標有折半信度、Cronbach's α 係數以及庫李(Kuder-Richardson) 20 號或 21 號公式。

此節採用 Winsteps 3.68.2 版所輸出之庫李 20 號公式之係數作為內部一致性

信度指標，描述 2013 年度兒童測驗正式考試信度。

表 6 顯示萌芽級聽力理解測驗與閱讀理解測驗信度係數分別為.92 及.67，閱讀理解測驗內部一致性信度低於聽力理解測驗，推測可能原因為，閱讀理解測驗試題數為 15 題，在測驗長度(test length)較聽力測驗短的情況下，導致內部一致性信度較低。

表 6 萌芽級內部一致性信度摘要表

測驗內容	信度係數
聽力理解	.92
閱讀理解	.67

而在成長級測驗以及茁壯級測驗中，聽力理解測驗與閱讀理解測驗的題長皆相同，因此，由表 7 與表 8 可知，成長級和茁壯級聽力理解測驗與閱讀理解測驗的內部一致性信度差異不大。

表 7 成長級內部一致性信度摘要表

測驗內容	信度係數
聽力理解	.84
閱讀理解	.83

表 8 茁壯級內部一致性信度摘要表

測驗內容	信度係數
聽力理解	.84
閱讀理解	.85

綜上所述，各等級聽力理解測驗信度係數介於.84 至.92 之間，閱讀理解測驗信度係數則介於.67 至.85 之間，除萌芽級閱讀測驗信度外，此數據顯示 2013 年兒童華語文聽力與閱讀理解正式考試卷信度大致良好。

(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)。由於無法直接觀察目標測量能力，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。內容效度的相關證據主要為評估、分析測驗所測量的能力以及各內容題數比重分配，是否符合測驗所欲測量的能力定義，通常會請所欲測量能力之學科專家評估試題是否符合上述要求，經由專家提供試題內容效度的方式則稱為專家效度；建構效度相關證據指的是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同，試題分析以及因素分析常用來提供建構效度的證據；受測者進行測驗的同時，調查受測者自評使用中文與家人交談的頻率，分析其使用頻率與測驗結果的相關程度，屬於同時效度，可做為效標效度的一種證據。

以下將從專家效度、試題分析、因素分析之驗證性因素分析(confirmatory factor analysis)，與同時效度等四方面來描述 2013 年度兒童測驗之內容效度、建構效度以及效標效度。

1. 內容效度

本會發展兒童測驗之初，集結華語教學、語言學、兒童心理與教育學相關領域專家共同擬訂各等級測驗目標，接著依據各等級目標訂定雙向細目表(見頁 14-15，表 9 至 10)，命題教師即根據雙向細目表及相關文件命題。兒童測驗研發人員回收各等級試題後，經測驗標準化流程(如圖 1)中試題修審步驟之第二階段：專家學者外審，聘請語言教學專家與華語教師審查各等級試題是否符合雙向細目表設定的內容，若符合，則表示測驗試題大致測量到測驗目標，此專家學者外審階段提供了兒童華語文能力測驗專家效度。

表 9 兒童華語文聽力測驗雙向細目表

題型 / 測驗等級 / 能力描述	聽句子連連看	聽對話連連看	看圖回答		聽單句選圖	聽對話選圖		完成對話			會話理解	段落理解
	萌芽	成長	萌芽	成長	萌芽	成長	茁壯	萌芽	成長	茁壯	茁壯	茁壯
能聽懂單句。	5				7							
能聽懂問句。			6									
能聽懂簡短對話並掌握關鍵訊息。		7				6						
能聽懂有關日常生活的問題。				6								
針對日常生活的話題，能理解並使用熟悉的表達方式以及非常基本的短語來應付所需。								7	6	7		
能理解簡短對話的內容，並掌握對話的主旨。							7				8	
能在聆聽日常生活中常見並簡短的宣布、說明、和語音媒體訊息時，掌握其主旨及重點資訊。												8

表 10 兒童華語文閱讀測驗雙向細目表

題型 / 測驗等級 / 能力描述	看圖辨義	圖文連連看		選詞填空			閱讀材料		短文理解	
	萌芽	萌芽	成長	萌芽	成長	茁壯	成長	茁壯	成長	茁壯
理解句子中的關鍵詞彙。	5	4	4							
能理解非常簡短、簡單的句子，會基本的單詞及會使用少數的句型。				6						
能理解簡單短文，並能正確使用基本的詞彙和簡單的語法。					6					
能讀懂簡短材料中的短語。							6			
在視覺輔助之下，能理解簡單、簡短的故事和敘事短文。									9	
能理解簡單的篇章，並正確使用語法和詞彙。						10				
能理解具日常功能性的簡單閱讀材料，並掌握其中的主旨與關鍵資訊。								10		
能理解簡單的故事與敘事短文，理解主要情節。										10

2. 建構效度

(1) 試題分析

兒童測驗組卷方式是依據 IRT 而來。IRT 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測量相同潛在特質，若忽略受測者回答試題是否仰賴單一特質這項資訊時，所獲得的試題參數及受測者能力估計值將會存在偏誤。

本節將採用 Winsteps 3.68.2 版的 Rasch 模式(如公式 1 所示)分析資料，以提供兒童測驗建構效度證據，結果如下所述。

由表 11 與表 12 可發現，萌芽級聽力理解測驗，試題的平均難度與考生能力估計值分別為-0.31 與 1.09，兩者差距 1.4 logits；閱讀測驗方面，試題的平均難度與受測者能力估計值則分別為-0.21 與 0.41，兩者差距 0.6 logit。顯示萌芽級測驗中，相較於聽力理解測驗，閱讀測驗之受測者平均能力與試題平均難度之間絕對值差異較小。

表 11 萌芽級試題難度估計分布

測驗內容	聽力	閱讀
平均值	-0.31	-0.21
標準差	0.83	0.57
最大值	1.25	1.26
最小值	-1.96	-1.20

表 12 萌芽級受測者能力估計分布

測驗內容	聽力	閱讀
平均值	1.09	0.41
標準差	2.13	1.20
最大值	4.44	3.86
最小值	-3.01	-4.24

成長級聽力與閱讀測驗結果(如表 13 與表 14)類似於萌芽級，試題的平均難度估計值與受測者能力估計值分別為聽力-0.21 與 3.27 以及閱讀-0.44 與 1.26，兩者差距各約為 3.5 logits 與 1.7 logits，閱讀測驗之受測者平均能力與試題平均難度之間絕對值差異較小。

表 13 成長級試題難度估計分布

測驗內容	聽力	閱讀
平均值	-0.21	-0.44
標準差	1.17	0.62
最大值	1.84	1.21
最小值	-3.23	-1.42

表 14 成長級受測者能力估計分布

測驗內容	聽力	閱讀
平均值	3.27	1.26
標準差	1.49	1.47
最大值	4.76	4.18
最小值	-0.92	-5.02

同樣的結果，在茁壯級(如表 15 及表 16)也可發現，聽力理解測驗中受測者平均能力與試題平均難度之絕對值差異比閱讀理解測驗大，兩者分別約為 3.0 logits 與 1.0 logits。

表 15 茁壯級試題難度估計分布

測驗內容	聽力	閱讀
平均值	-0.19	0.01
標準差	0.74	0.60
最大值	1.41	1.24
最小值	-2.46	-1.65

表 16 茁壯級受測者能力估計分布

測驗內容	聽力	閱讀
平均值	2.83	1.06
標準差	1.34	1.40
最大值	4.64	4.77
最小值	-1.50	-1.27

整體來說，兒童測驗本年度聽力與閱讀測驗各等級的考生能力估計值皆高於試題難度的平均值。造成此一現象可能的原因為，本測驗萌芽級、成長級、茁壯級的建議報考學時分別為 150 小時、300 小時與 450 小時；而萌芽級考生中超過建議報考時數的人數達有效填答樣本的四成，至於成長級和茁壯級有效填答樣本

中，學習時數超過 600 小時的比例分別為 60.7%與 84.7%。由此可見，有相當高比例的考生學習時數遠高於建議報考時數，可能因此使得整體考生的能力值高於所報考等級的平均難度。

此外，各等級閱讀測驗受測者平均能力與試題平均難度之間絕對值差異比聽力測驗小，造成此現象的可能原因為參加本測驗的兒童多屬華裔，當家庭成員同屬華裔時，便可能常以華語交談，因此學習者在學習過程中可大量依靠聽和說的方式學習華語，此一現象反映於受測者聽力平均能力在三個等級皆有偏高的趨勢(見表 12、表 14 及表 16)；相較之下，在非華語地區，學習者大多數僅能透過學校課程及課外讀物學習中文字詞，能接觸到的視覺接收活動較少，因此，此一現象反映於受測者的閱讀平均能力與試題平均難度差異較小。同樣的現象可由圖 2 至圖 4 受測者與試題分布圖(Person – Item Map)觀察得知，將受測者能力估計值與試題難度估計值分別置於圖的左側和右側，由上而下將受測者能力由高至低排列、試題難度由難至易排列，可發現與聽力理解測驗相較，閱讀理解測驗試題的難度涵蓋範圍與受測者能力分布範圍間的差距較小，各等級聽力與閱讀理解測驗試題平均難度 95%信賴區間涵蓋的受測者能力分布分別為，萌芽級 61%與 72%、成長級 23%與 37%以及茁壯級 12%與 65%。

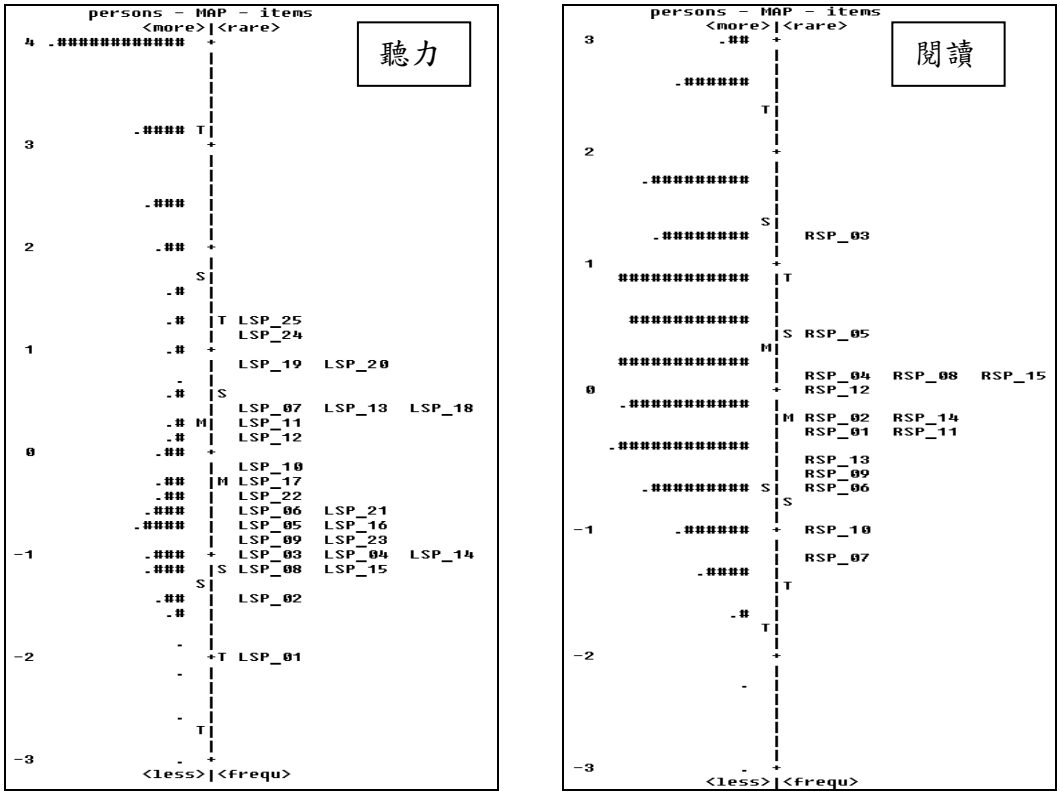

圖 2 萌芽級測驗受測者與試題分布圖

註:“#”在聽力與閱讀理解測驗分別表示 19 位及 11 位受測者;“.”表示一位受測者;LSP_01~LSP_25 和 RSP_01~RSP_15 表示試題編號。

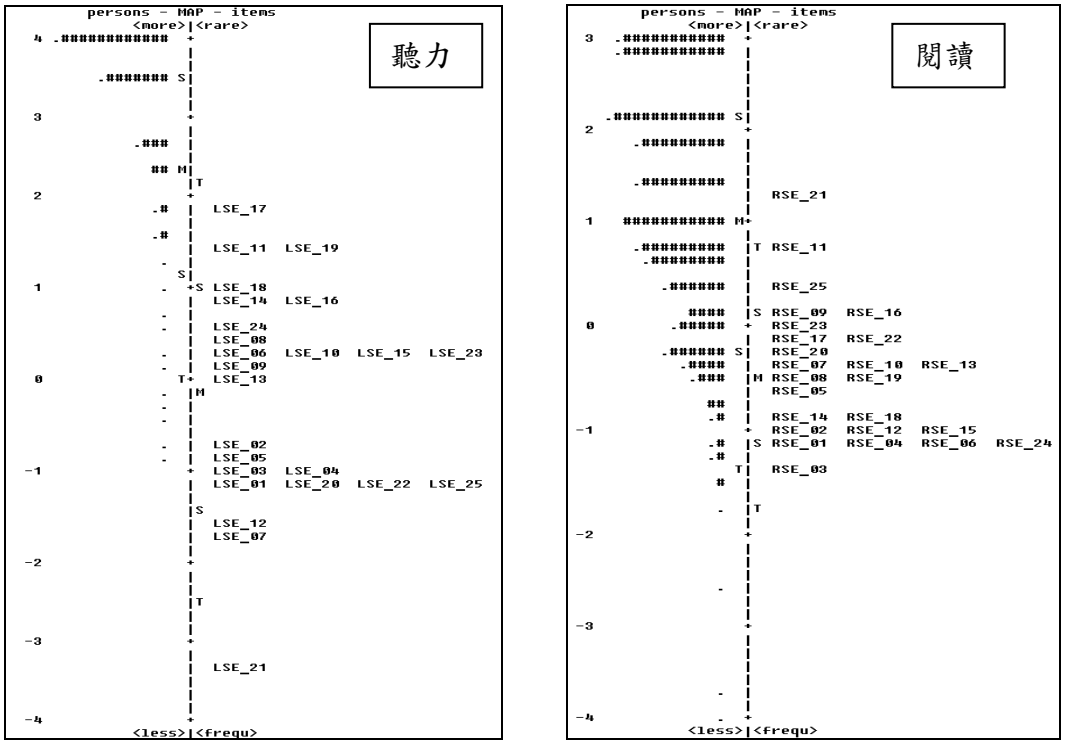

圖 3 成長級測驗受測者與試題分布圖

註:“#”在聽力與閱讀理解測驗分別表示 22 位及 6 位受測者;“.”表示一位受測者;LSE_01~LSE_25 和 RSE_01~RSE_25 表示試題編號。

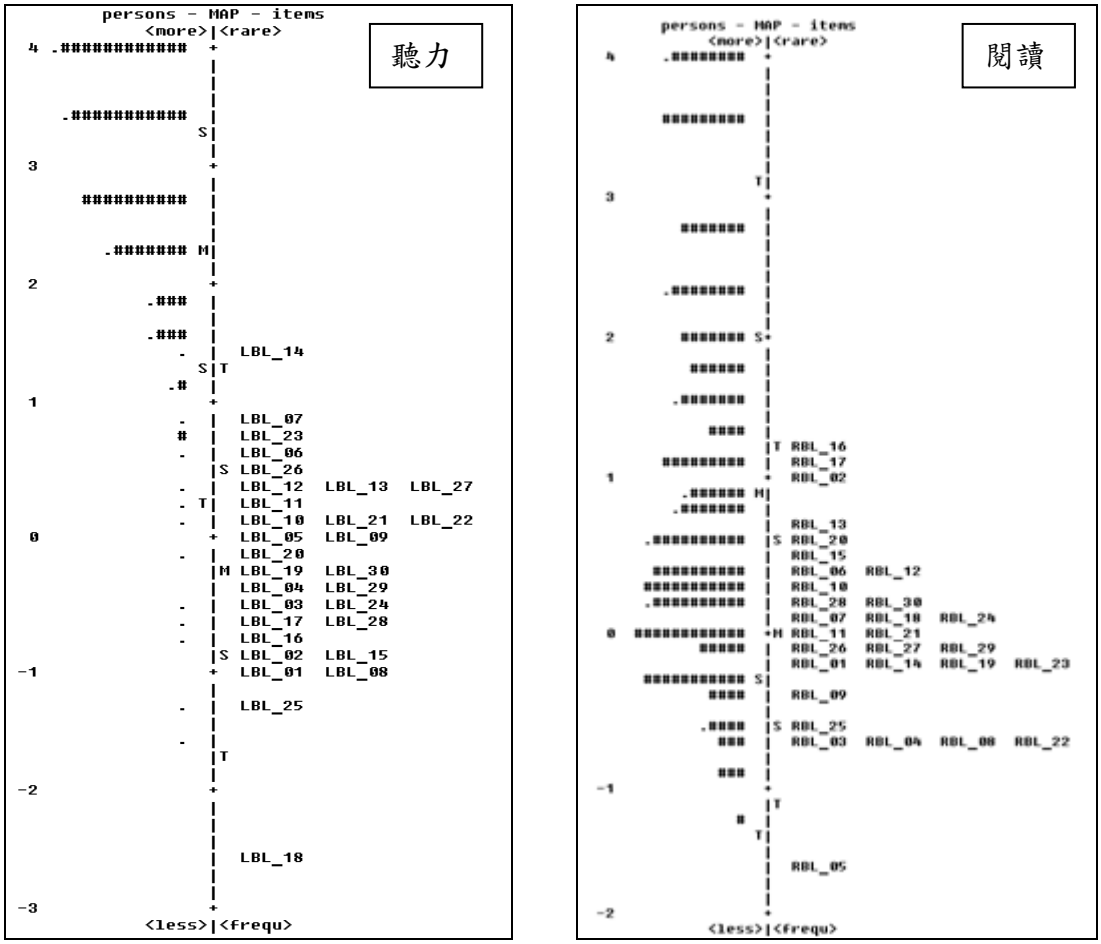

圖 4 茁壯級測驗受測者與試題分布圖

註：“#”在聽力與閱讀理解測驗分別表示 6 位及 2 位受測者；“.”表示一位受測者；LBL_01~LBL_30 和 RBL_01~RBL_30 表示試題編號。

在試題品質分析方面，採用統計指標 Infit MNSQ 介於 0.6 到 1.4 以及 ZSTD 介於-3 到 3 的標準評估試題與單向度 IRT 模式適配的程度。結果如表 17 所示，萌芽級聽力測驗 25 道試題中有三道、萌芽級閱讀測驗 15 道試題中有一道，成長級聽力測驗 25 道試題中有六道、成長級閱讀測驗 25 道試題中有三道、以及茁壯級聽力測驗 30 道試題中有十道試題，茁壯級閱讀測驗 25 道試題中有二道不符合單向度 IRT 模式的假設。各等級在聽力與閱讀測驗適配率分別為，萌芽級 88%與 93%、成長級 76%與 88%以及茁壯級 67%與 93%。

一般來說，造成試題與模式不適配的原因有以下幾種，一為當受測者答題所需要的能力超過單一向度時，則可能造成試題不適配，因在 Rasch 模式的假設下，受測者答對試題機率僅受到單向度能力(此為兒童聽力或閱讀能力)所影響。二是當試題具有其他影響答題機率的特性，例如，試題的猜測力或鑑別度，而未能被 Rasch 模式納入解釋時，也會發生試題不適配情形。三是當受測者因為粗心

或其他原因使得作答未符合其能力表現時，亦會造成試題估計不符合模式。例如，在模式假定受測者能力越高答對機率越高的條件下，高能力之受測者卻答錯較簡單題目。

由於本測驗正式卷的試題皆經過預試，且獲得穩定難度參數估計值後所組成，故測驗觀察資料與測量模式不適配的原因應為第三種情形；在分析時，試題難度被設定為已知再進行模式適配度考驗，而受測者因粗心或其他原因使得作答未符合其能力表現，造成試題估計不符合模式。

表 17 試題適配分布

測驗等級	萌芽級		成長級		茁壯級	
正式考試	聽力	閱讀	聽力	閱讀	聽力	閱讀
總試題數	25	15	25	25	30	30
適配題數	22	14	19	22	20	28
不適配題數	3	1	6	3	10	2
試題適配率 (%)	88	93	76	88	67	93

以下進一步針對各等級測驗不適配試題進行選項分析，結果如表 18 至表 20 所示。表 18 至表 20 列出各等級測驗中不適配試題之難度參數估計值及估計誤差、各選項選答人數、選答該選項的受測者平均能力估計值、平均能力估計誤差及點二系列相關係數。由表 18 可發現，萌芽級聽力測驗不適配試題中，選答正確選項人數都是最多的，除了 LSP_20 外，正答選項平均能力估計值也是最高的，表示選答該選項的受測者平均能力高於選答錯誤選項者，如模式所預期，能力越高的受測者答對試題的機率越高。所有試題中正確選項的點二系列值都較高且皆為正值(介於.39 至.51 之間)，表示具有幅合效度(convergent validity)，即能力越高者答對試題的機率越高；錯誤選項值較低且幾乎為負值，表示具有區辨效度(discriminate validity)，可有效區分能力高與低的受測者(Linacre，2010)。LSP_20 正答考生平均能力值雖略低於複選考生，但複選人數僅有 1 人，而其他誘答選項考生平均能力值均低於正答，加上此題其他選項分析結果皆無異常，此題仍具有效度。

閱讀測驗方面，RSP_03 選正答(O)與誘答(X)的人數接近，考生平均能力值為正答較高，但正答點二系列相關結果為.04，低於.20。由圖 5 試題特徵曲線(item characteristic curve，簡稱 ICC)來看，能力值低於此題難度參數(1.257)約 0.5 至

3.0，及 4.0 至 4.5 logits 的考生，答對比例高於模式預期，使得相關係數較低，也造成觀察資料與模式不適配。如：能力值低於難度參數 1.5 logits 的考生，依照模式預期，答對率應約為 0.18，但實際作答資料顯示，此一能力值考生答對率約在 0.3。

表 18 萌芽級測驗不適配試題選項描述性摘要表

測驗類別	試題編號	難度	估計誤差	選項	計分	人數	平均能力	平均能力估計誤差	點二系列相關
聽力	LSP_01	-1.958	0.094	B	0	218	-0.79	0.05	-.43
				*	0	1	-0.78	--	-.03
				A	0	68	-0.72	0.10	-.21
				C	1	881	1.75	0.07	.51
				M	0	29	-0.71	0.26	-.13
	LSP_20	0.898	0.082	C	0	143	-0.67	0.09	-.30
				A	0	193	-0.32	0.08	-.24
				*	0	1	1.94	--	.03
				B	1	855	1.71*	0.07	.41
				M	0	5	-1.19	0.15	-.08
	LSP_25	1.246	0.086	B	0	13	-1.15	0.09	-.12
				D	0	129	-0.82	0.05	-.32
				C	0	12	-0.74	0.21	-.09
				F	0	25	-0.65	0.13	-.12
				*	0	1	-0.59	--	-.02
				E	0	12	0.00	0.34	-.04
				A	1	976	1.53	0.07	.39
				M	0	29	-1.39	0.12	-.20
閱讀	RSP_03	1.257	0.070	X	0	563	0.21	0.04	-.04
				O	1	566	0.75	0.05	.04
				M	0	68	-0.73	0.15	-.20

註：M 表示漏答；“*”表示複選；“*”表誘答選項考生平均能力值高於正答；“--”表人數為 1 人，無法計算平均能力估計標準差。

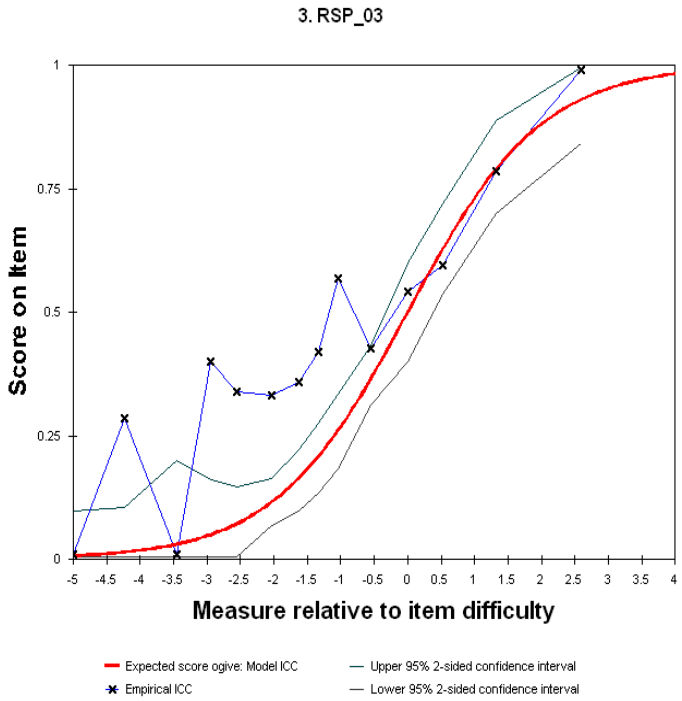

圖 5　RSP_03 試題特徵曲線

表 19 為成長級聽力、閱讀測驗選項分析的結果，與萌芽級相似，不適配試題中選答正確選項人數都是最多的，該選項平均能力估計值也是最高的，正確選項的點二系列值介於.17 至.47 之間，說明成長級中，在數據上呈現為不適配的這幾道試題仍具輻合效度與區辨效度，可有效區分能力高與低的受測者(Linacre，2010)。

表 19 成長級測驗不適配試題選項描述性摘要表

測驗類別	試題編號	難度	估計誤差	選項	計分	人數	平均能力	平均能力估計誤差	點二系列相關
聽力	LSE_01	-1.184	0.223	C	0	14	1.11	0.40	-.20
				B	0	33	2.15	0.22	-.07
				A	1	674	3.37	0.06	.17
				M	0	2	1.55	0.31	-.03
	LSE_02	-0.740	0.191	B	0	6	-0.2	0.39	-.28
				A	0	2	0.13	0.22	-.13
				C	1	715	3.31	0.06	.31
	LSE_05	-0.848	0.198	C	0	3	-0.08	0.24	-.19
				A	0	15	0.14	0.17	-.37
				B	1	703	3.36	0.05	.42
				M	0	2	0.24	0.53	-.13
	LSE_20	-1.083	0.215	B	0	1	-0.49	--	-.13
				D	0	1	0.13	--	-.09
				E	0	2	0.23	0.10	-.12
				*	0	1	0.77	--	-.06
				C	1	716	3.29	0.06	.20
				M	0	2	2.47	0.25	.01
	LSE_22	-1.205	0.225	C	0	2	-0.18	0.31	-.16
				*	0	1	0.13	--	-.09
				D	0	2	0.45	0.32	-.11
				A	0	2	1.00	0.00	-.06
				E	0	1	1.53	--	-.02
				B	1	713	3.31	0.06	.20
				M	0	2	1.74	0.48	-.02
	LSE_25	-1.198	0.224	A	0	25	0.76	0.19	-.36
				D	0	18	1.01	0.22	-.25
				B	0	28	1.46	0.15	-.18
				*	0	1	2.22	--	.00
				C	1	642	3.55	0.05	.47
				M	0	9	0.84	0.48	-.21
閱讀	RSE_05	-0.580	0.103	A	0	196	0.57	0.07	-.22
				B	1	497	1.64	0.07	.22
				M	0	30	-0.46	0.26	-.26
	RSE_12	-0.959	0.112	A	0	100	0.15	0.09	-.28
				C	0	93	0.38	0.08	-.17
				B	1	526	1.65	0.06	.35
				M	0	4	-2.52	1.12	-.21
	RSE_24	-1.064	0.115	X	0	189	0.26	0.07	-.38
				O	1	519	1.68	0.06	.38
				M	0	15	-0.51	0.45	-.17

註：M 表示漏答；“*”表示複選；“--”表人數為 1 人，無法計算平均能力估計標準差。

表 20 為茁壯級測驗的選項分析結果，所有不適配試題選答正確選項人數都是最多的，該選項平均能力估計值也是最高的，除了 LBL_08 外，其餘試題正確選項的點二系列值最高且並為正值(.20 至.56)，表示試題具有輻合效度；錯誤選項的點二系列值較低且幾乎為負值，表示試題具有區辨效度，可有效區分能力高與低的受測者(Linacre，2010)。

由圖 6 LBL_08 的試題特徵曲線來看，能力值高於此題難度參數(-0.960)約 2.5 至 5.0 logits 的考生，答對比例低於模式預期，使得相關係數較低，也造成觀察資料與模式不適配。如：能力值高於難度參數 2.5 logits 的考生，依照模式預期，答對率應約為 0.92，但實際作答資料顯示，此一能力值考生答對率低於 0.75。

表 20 茁壯級測驗不適配試題選項描述性摘要表

測驗類別	試題編號	難度	估計誤差	選項	計分	人數	平均能力	平均能力估計誤差	點二系列相關
聽力	LBL_05	-0.045	0.208	A	0	7	-0.15	0.38	-.45
				B	0	13	0.86	0.17	-.28
				C	1	312	2.98	0.07	.50
	LBL_06	0.598	0.175	A	0	6	0.26	0.45	-.33
				B	0	7	0.44	0.39	-.31
				C	1	319	2.93	0.07	.45
	LBL_08	-0.960	0.281	B	0	8	1.52	0.52	-.13
				A	0	38	2.29	0.14	.03
				C	1	286	2.94	0.08	.03
	LBL_10	0.102	0.199	A	0	22	1.59	0.23	-.18
				B	0	37	1.63	0.19	-.23
				C	1	273	3.09	0.07	.31
	LBL_12	0.383	0.185	B	0	20	1.28	0.31	-.29
				A	0	40	1.87	0.15	-.12
				C	1	272	3.08	0.08	.28
	LBL_15	-0.875	0.273	C	0	2	-0.99	0.51	-.35
				B	0	6	0.16	0.52	-.35
				A	1	324	2.90	0.07	.48
	LBL_17	-0.569	0.246	B	0	5	-0.73	0.34	-.51
				C	0	8	0.70	0.39	-.27
				A	1	319	2.94	0.07	.53
	LBL_19	-0.218	0.219	A	0	20	1.55	0.31	-.21
				B	0	45	2.01	0.15	-.09
				C	1	267	3.06	0.08	.21
	LBL_26	0.514	0.178	A	0	17	1.47	0.27	-.19
				C	0	74	1.89	0.13	-.21
				B	1	241	3.21	0.08	.29
	LBL_28	-0.636	0.251	A	0	1	-1.31	--	-.28
				C	0	6	-0.43	0.48	-.48
				B	1	325	2.90	0.07	.56

表 21 茁壯級測驗不適配試題選項描述性摘要表(續)

測驗類別	試題編號	難度	估計誤差	選項	計分	人數	平均能力	平均能力估計誤差	點二系列相關
閱讀	RBL_05	-1.653	0.188	B	0	134	0.53	0.07	-.20
				A	1	194	1.44	0.11	.20
				M	0	4	0.22	0.71	-.07
	RBL_08	-0.722	0.145	B	0	114	0.38	0.09	-.30
				A	1	217	1.42	0.10	.30
				M	0	1	-0.28	--	-.06

註：M 表示漏答；“*”表示複選；“--”表人數為 1 人，無法計算平均能力估計標準差。

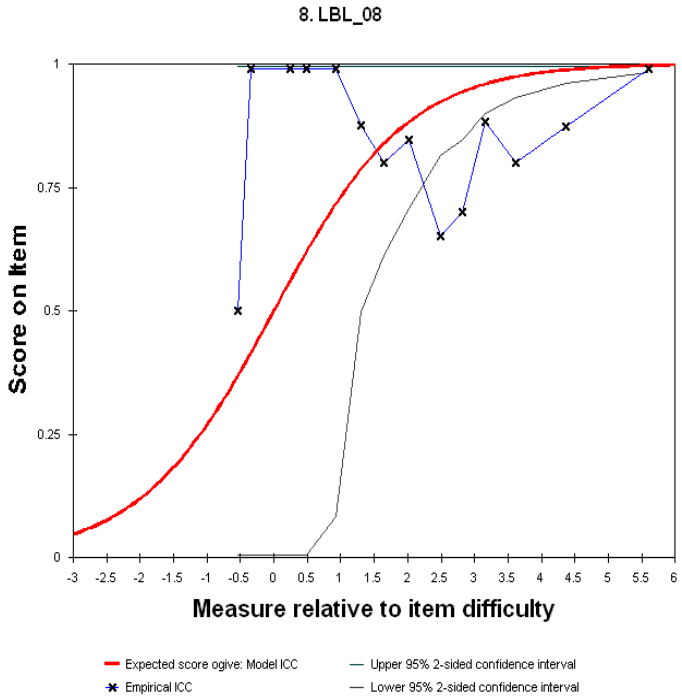

圖 6 LBL_08 試題特徵曲線

(2) 驗證性因素分析

除了透過試題分析來評估兒童測驗是否具有建構效度之外，本報告亦從「結構方程模型(structural equation modeling)驗證性因素分析」來評估測驗的建構效度。由於兒童測驗題型雖包含選擇題、是非題以及連連看題型，然而，答題反應皆為非對即錯的二分名義變項，不宜直接將資料使用於因素分析，因此，先計算測驗中各題型總分，以題型總分做為測量變項評估是否各等級測驗之不同題型可組合為單維(uni-dimensionality)能力(此指兒童測驗聽力理解能力或閱讀理解能力)。故每一種因素分析模型中，題型表示測量變數，而欲測得能力(聽力或閱讀理解能力)表示潛在變數，例如，成長級聽力測驗中，測量變數為四種題型(如表

2)，聽力理解能力為潛在變數。

此節使用Mplus 7.0版進行資料分析，估計方法採用SEM最常用的參數估計法「最大概似法」(maximum likelihood，ML)，各等級測驗之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(1998)，以及Hu和Bentler(1998)的研究結果，並綜合本研究觀察變項屬性，與軟體輸出報表所提供指標，在基本適配度部分的評估標準如下：

(1) 因素負荷量宜介於.50至.95之間。

(2) 不能有過大的標準誤。

(3) 誤差變異數不得為負。

(4) 誤差變異達到顯著。

至於整體適配指標部分，則採用卡方考驗(χ^2)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)指標與標準化殘差均方根指標(standardized root mean square residual；簡稱SRMR)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不顯著(p>.05)、RMSEA<.08、SRMR<.08、CFI和NNFI >.90。

在基本適配指標部分，各等級聽力與閱讀單因素驗證性因素分析結果如圖 7 至 12 所示。萌芽級聽力測驗單因素模型驗證性因素分析結果顯示，因素負荷量介於.77 至.87 間；因素負荷量標準誤約為.01；因素負荷量皆達顯著水準(p<.05)；誤差變異量皆為正值且達顯著水準(p<.05)。萌芽級閱讀測驗單因素模型驗證性因素分析結果為，因素負荷量介於.39 至.68 間；各因素負荷量標準誤介於.04 至.06；因素負荷量皆達顯著水準(p<.05)；誤差變異量為正值且皆達顯著水準(p<.05)。

根據以上指標分析結果，可知萌芽級聽力測驗與閱讀測驗單因素驗證性因素模型基本適配度檢驗大致良好，除了看圖辨別句義、選詞填空略低外(.39 與.44)，各題型因素負荷量均介於.50 至.95 之間。

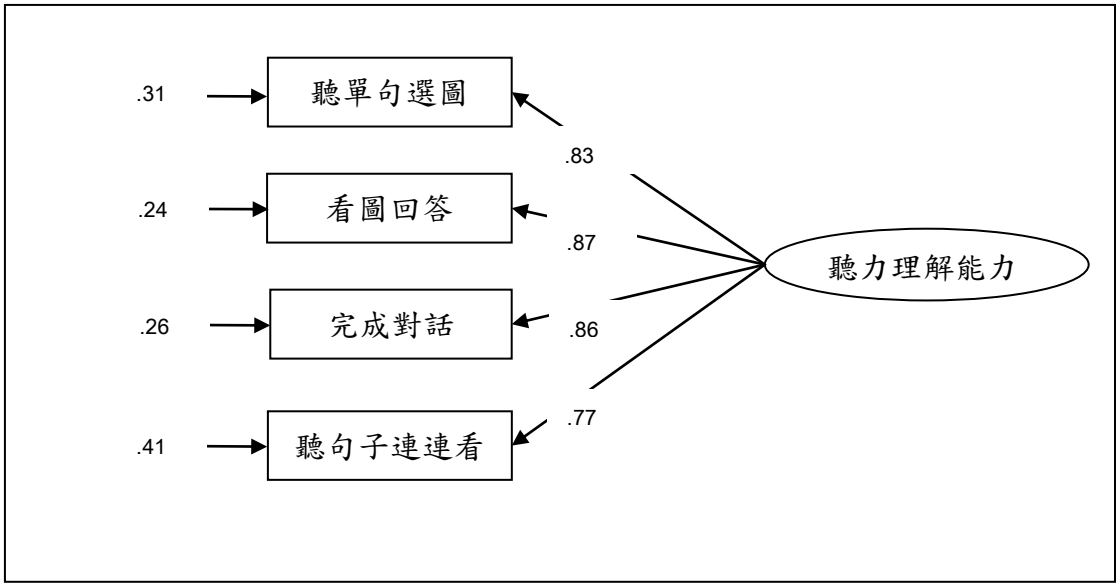

圖 7 萌芽級聽力測驗單因素驗證性因素分析

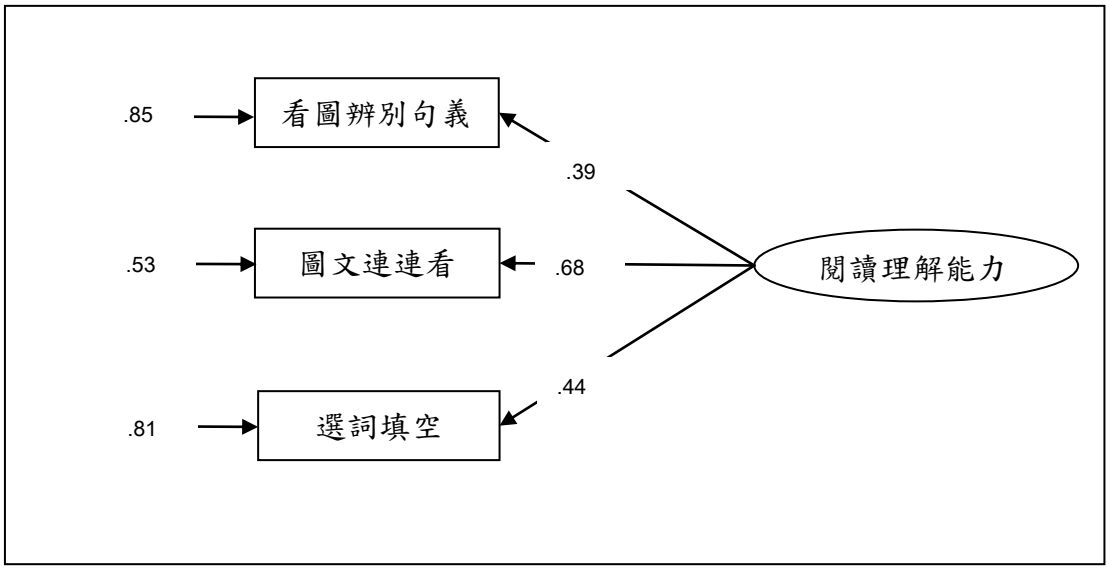

圖 8 萌芽級閱讀測驗單因素驗證性因素分析

成長級聽力測驗與閱讀測驗單因素模型驗證性因素分析顯示(如圖 9 與 10)：聽力測驗部分，因素負荷量介於.71 至.78 之間；因素負荷量標準誤約為.02；因素負荷量達顯著水準(p<.05)；誤差變異量為正值且達顯著水準(p<.05)。閱讀測驗部分，因素負荷量介於.53 至.75 之間；因素負荷量標準誤約為.03；因素負荷量達顯著水準(p<.05)；沒有負值的誤差變異量且皆達顯著水準(p<.05)。以上分析顯示，成長級聽力測驗與閱讀測驗單因素模型基本適配度檢驗良好，各題型因素負荷量皆介於.50 至.95 之間。

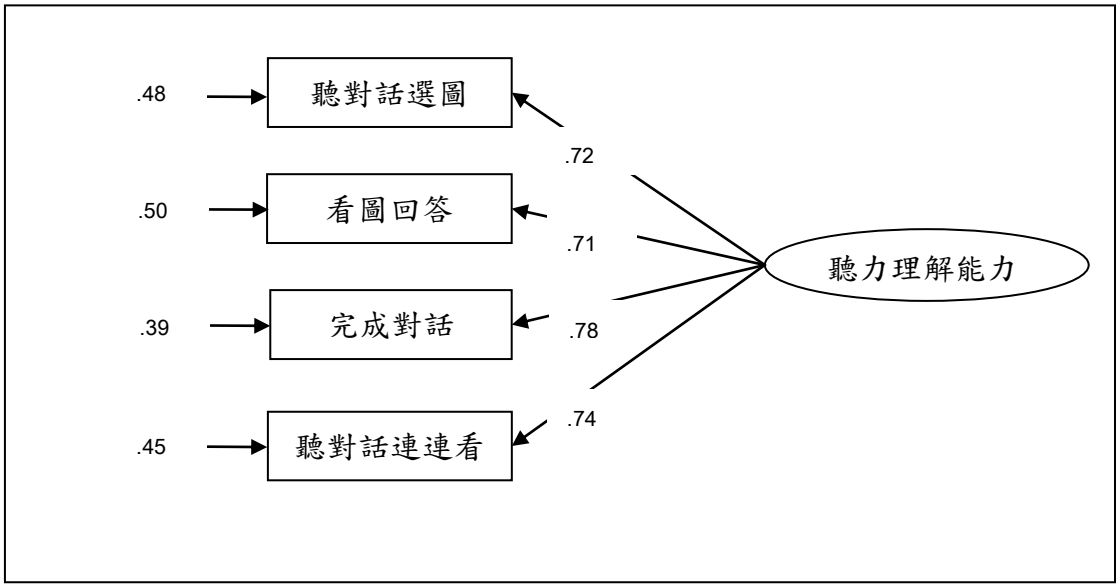

圖 9 成長級聽力測驗單因素驗證性因素分析

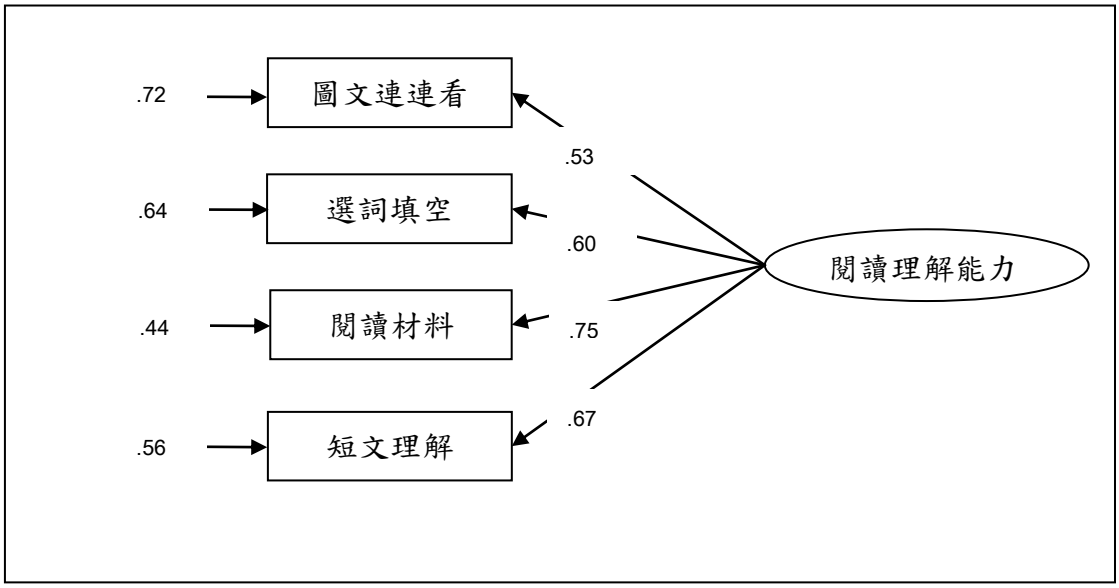

圖 10 成長級閱讀測驗單因素驗證性因素分析

茁壯級聽力測驗與閱讀測驗驗證性因素分析結果如圖 11 和 12 所示，聽力單因素模型顯示，因素負荷量為.60 至.84 之間；其標準誤介於.03 至.04 之間；因素負荷量達顯著水準(p<.05)；誤差變異量為正值且皆達顯著水準(p<.05)。閱讀測驗因素負荷量則是閱讀材料較低，數值為.30，其餘兩個題型因素負荷量分別為.70 及.80；因素負荷量之標準誤介於.06 至.09 之間；因素負荷量達顯著水準(p<.05)；誤差變異量介於.36 到.91 之間且皆達顯著水準(p<.05)。以上分析顯示茁壯級聽力測驗與閱讀測驗單因素模型基本適配度檢驗大致良好，除閱讀材料略低外(.30)，其餘各題型因素負荷量皆介於.50 至.95 之間。

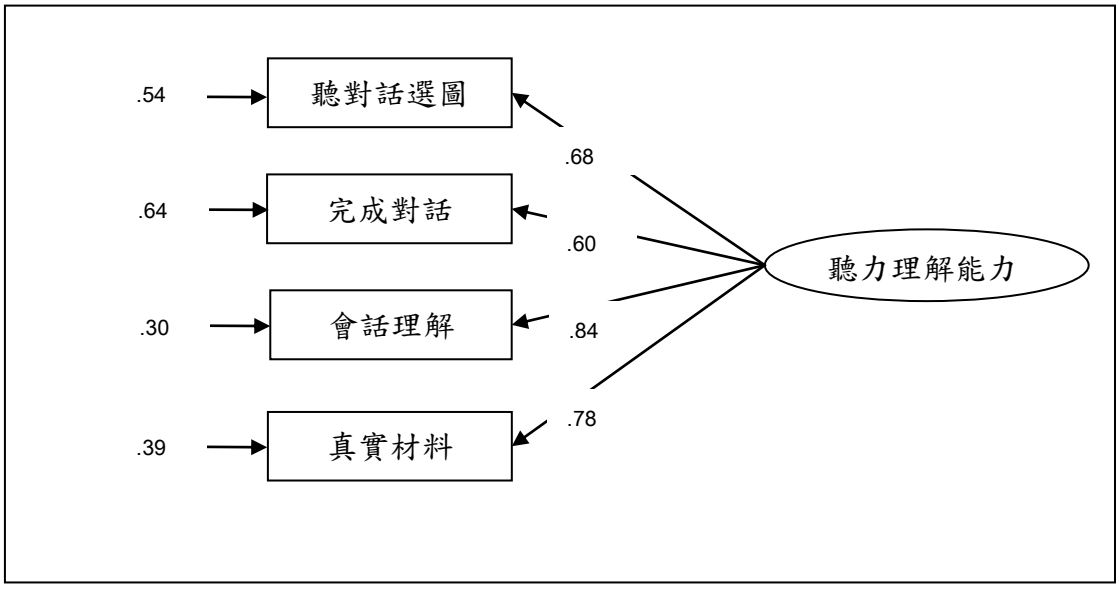

圖 11 茁壯級聽力測驗單因素驗證性因素分析

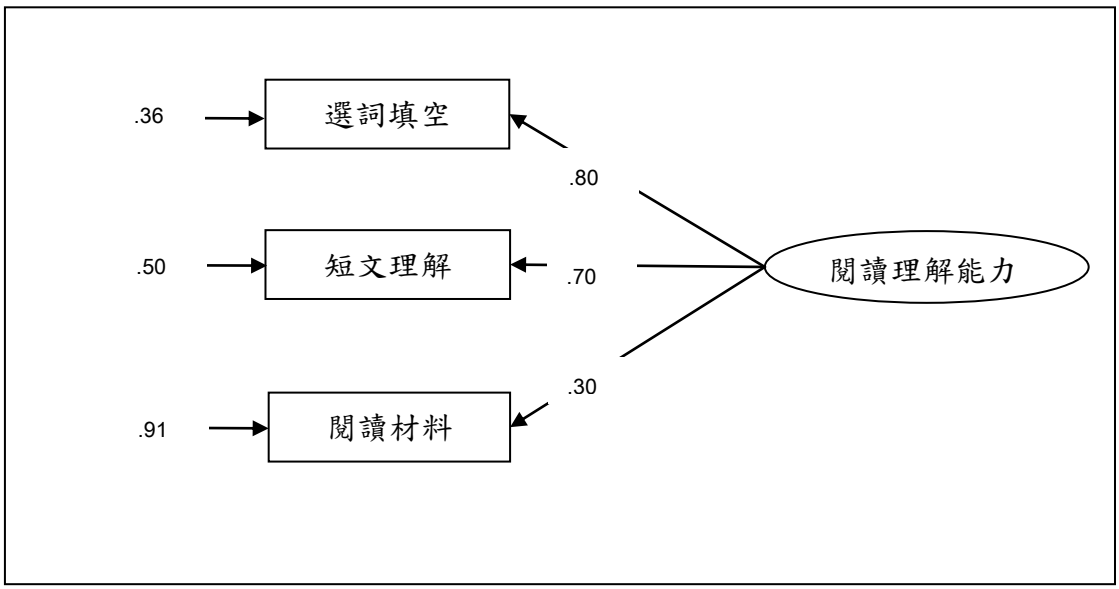

圖 12 茁壯級閱讀測驗單因素驗證性因素分析

整體模式適配度主要在評鑑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，經由卡方考驗初步檢驗整體模式適配度(如表 21 所示)，萌芽級、成長級和茁壯級聽力測驗三個模型的 χ^2 值分別為 1.833、4.292 和 1.382，皆未達顯著水準(p>.05)，顯示此三測量模型與觀察資料適配；成長級閱讀測驗單因素模式的 χ^2 值為 6.158，雖達顯著水準(p<.05)，但數值相當接近臨界點(p =.046)，顯示此模型與觀察資料僅略為不同；萌芽級閱讀測驗與茁壯級閱讀測驗二模型的 χ^2 值皆為 0，是因為兩模式的測量資料數與參數估計數相等，即此二模式與觀察資料呈現完美適配。

在整體適配指標部分，各項指標結果如表 21 所示。由於萌芽級與茁壯級閱讀測驗單因素模式顯示與觀察資料完美適配，因此，軟體輸出報表中沒有適配指標，無法進行模型適配度考驗，故此部分僅對萌芽級聽力測驗、成長級聽力測驗、成長級閱讀測驗以及茁壯級聽力測驗單因素模型進行適配指標考驗。在絕對適配度評估上，結果顯示，萌芽級、成長級與茁壯及聽力測驗單因素模式，和成長級閱讀測驗單因素模式，其 RMSEA 和 SRMR 數值皆符合判斷標準，小於.08，四種模式都符合絕對適配度指標；而在增值適配度方面，上述四種模型之 CFI 和 NNFI 兩項指標數值皆大於.90，符合增值適配度指標。

綜合上述可得出以下結論，兒童測驗各等級測驗聽力測驗與閱讀測驗具有建構效度，各測驗題型分別可測得聽力理解能力或閱讀理解能力。

表 21 模式適配度卡方統計量與整體模式適配度指標摘要表

檢驗模型	卡方檢驗	絕對適配度		增值適配度	
	χ^2	RMSEA	SRMR	CFI	NNFI
萌芽級聽力單因素模式	1.833	.000	.003	1.00	1.00
萌芽級閱讀單因素模式	0	-	-	-	-
成長級聽力單因素模式	4.292	.040	.009	1.00	.99
成長級閱讀單因素模式	6.158*	.054	.016	.99	.98
茁壯級聽力單因素模式	1.382	.000	.008	1.00	1.00
茁壯級閱讀單因素模式	0	-	-	-	-

註: *表示 $p < .05$；χ^2 為 0 表示該測驗單因素驗證性因素模型之測量變數為三個，測量資料數與自由估計參數個數相等，所估計的模型稱為飽和模型(saturated model)，卡方統計量為 0，呈現估計模型與實際模型完美適配。

3. 效標效度

表 22 為考生使用中文與家人交談的頻率與萌芽級、成長級、茁壯級聽力測驗和閱讀測驗得分間之關係。獨立樣本 t 檢定結果發現，使用中文交談頻率不同之考生，除在成長級閱讀測驗分項未達顯著外，其他各等級或各類測驗分數均達顯著差異(p<.05 或 p<.01)；「經常」使用中文與家人交談的考生，在聽力測驗、閱讀測驗的得分上均優於「非經常」使用中文與家人交談的考生。也就是說，受測者自評其是否經常使用中文與家人交談，可以作為預測受測者測驗成績的指標。

未達顯著的成長級閱讀測驗部分，填答「經常」與「非經常」使用中文與家人交談的考生平均分數差距甚小，將在未來測驗中持續觀察與家人以中文交談頻率一項，對不同等級和能力分項得分的影響和差異。

表 22 兒童測驗使用中文交談頻率不同考生之成績表現

測驗等級	測驗類別	中文交談頻率	人數	平均分數	*t* 值
萌芽級	聽力測驗	(1)經常	95	24.1	34.166**
		(2)非經常	708	12.9	
	閱讀測驗	(1)經常	95	11.2	9.693**
		(2)非經常	708	8.2	
成長級	聽力測驗	(1)經常	78	24.0	5.914**
		(2)非經常	278	22.1	
	閱讀測驗	(1)經常	78	19.2	-1.287
		(2)非經常	278	19.9	
茁壯級	聽力測驗	(1)經常	88	28.4	6.06**
		(2)非經常	103	25.2	
	閱讀測驗	(1)經常	88	20.3	1.990*
		(2)非經常	103	18.7	

註：*p<.05；**p<.01

五、 結論

本文為 2013 年兒童華語文能力測驗技術報告，闡述內容共包含兩個部分，第一部分分別針對兒童華語文能力測驗之能力描述、測驗題型題數、通過門檻等方面進行概述，並說明測驗研發、施測和成績公布之標準化流程。第二部分則為 2013 年度之兒童測驗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並確切地測量受測者的目標潛在能力。

在測驗信度分析方面，本文採內部一致性指標觀察測驗試題間之相關性，藉以確認整份測驗中的試題是否皆測量到相同潛在特質及其程度為何。在測驗效度分析部分，本測驗在組卷完成後，即通過專家審查、評估試題於不同等級中的適切性以確保試題品質，針對試題內容的審核步驟完整且明確，可賴以確保測驗之內容效度。在此應補充說明的是，由於本測驗試題內容的適切性主要仰賴專家審查，唯有通過審查並完成預試的試題方能使用於正式考試中施測，故專家審查程序是重要且不可忽略的內容效度證據。為了提供更嚴謹、全面及量化的專家效度證據，本會未來在邀請外審專家審查時，除目前所提供的各等級能力描述與各大題所測能力等資訊外，也考慮提供更具針對性的試題檢核項目資料，如設計審題問卷，請專家針對每一道試題進行詳細評分或判斷，通過此一辦法獲得專家對於試題評估的量化結果。

除了具備測驗之內容效度證據之外，在施測完成後，我們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，與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，我們可以說，兒童測驗有效標效度的證據。

總體而言，本文之信度、效度分析結果顯示，本會於 2013 年度所舉辦之全球性兒童測驗正式考試，所獲得之受測者成績及結果已相當可靠，並可確實測量到測驗研發初期所訂定之目標能力。

六、 文獻

多媒體英語學會(譯)(2007)。**歐洲共同語文參考架構**。高雄：和遠圖書資訊出版社。(Council of Europe, 2001)

陳怡靜、趙家璧(2012)。寓試於樂—兒童華語文能力測驗。**華文世界，*109***，26-32。

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

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

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

Linacre, J.M. (2009). Winsteps® (Version 3.68.2) [Computer Software]. Beaverton, Oregon: Winsteps.com.

Linacre, J. M. (2010). Winsteps® Rasch measurement computer program User's Guide. Beaverton, Oregon: Winsteps.com

McKay, P. (2006). *Assessing Yong Language Learners*. Cambridge: Cambridge University Press.

Muthén, L.K. and Muthén, B.O. (1998-2013). Mplus User's Guide. Seventh Edition. Los Angeles, CA: Muthén & Muthén.

Rasch, G. (1960). *Probabilistic models for some intelligence and attainment tests* Copenhagen: Institute of Educational Research. (Expanded edition, 1980. Chicago: The University of Chicago Press.)

書　　名：兒童華語文能力測驗技術報告—2013(5)
聽力、閱讀測驗信效度
出 版 者：國家華語測驗推動工作委員會
24449 新北市林口區仁愛路一段 2 號
886-2-7734-5638
印 刷 者：上校文化印刷有限公司
80744 高雄市三民區通化街 88 巷 26 號
886-7-311-6011
出版日期：2015 年 12 月
定　　價：新台幣 100 元