國立臺灣師範大學教育心理與輔導學系
教育心理學報，2013，45 卷，1 期，39-61 頁
DOI 10.6251/BEP.20121128

Rasch 模式建置國小高年級閱讀理解測驗*

林怡君　　張麗麗　　陸怡琮
高雄市立　　國立屏東教育大學
鼎金國民小學　　教育心理與輔導學系

「國小高年級閱讀理解測驗」係根據 NAEP 閱讀理解評量架構，以 Rasch 部份給分模式建置適用於高年級一般學童之文本閱讀理解測驗。研究對象係採分層隨機抽樣取得之高屏地區國小五、六年級學童 1,052 人。測驗共 39 題，含文學及訊息文本、選擇題及建構反應題，測「尋找和回憶」、「整合和解釋」及「批判和評鑑」三個閱讀理解認知層次。計量分析的結果顯示：Rasch 模式建置之測驗符合客觀測量特性（閱讀理解構念為單向度、具性別及年級恆等性）、建構反應題之等級間距適切且呈階層關係、試題難度分配廣且與學童能力相對應、閱讀理解認知層次具階層性、測驗具合理之聚斂與區辨相關且能區辨不同次群體（年級、性別、語文能力）、測驗具理想之分隔信度及評分者信度。本文亦針對測驗應用、Rasch 模式應用，以及後續研究提出建議。

關鍵詞：閱讀理解測驗、NAEP、Rasch 模式

二十一世紀是知識經濟的時代，擁有知識就有機會掌握經濟優勢，閱讀使得人們可以進入新的知識領域，有效掌握日新月異的知識。閱讀也是學生學習各學科的重要基礎。透過閱讀理解的能力，人們才能夠在這終身學習的時代持續自我成長。

有鑑於閱讀的重要性，教育部近年來也開始重視閱讀，譬如：2004 年針對弱勢地區國小推動「焦點三百--國小兒童閱讀計畫」、2006 年對偏遠地區國中小實施閱讀推廣計畫、2008 年推動「悅讀 101-教育部國民中小學閱讀提升計畫」等（教育部，2009）。在課程設計方面，九年一貫課程的國語文領域除了以「閱讀」取代「讀書」一詞外，更明訂「語文教學以閱讀為核心，兼顧聆聽、說話、作文、寫字等各項教學活動的密切聯繫」，以及「閱讀能力之評量，宜參考階段能力指標，檢覈其文字理解與詞語辨析、文意理解與大意摘取、統整要點與靈活應用、內容深究與審美感受等向度」（教育部，2008）。

* 本篇論文通訊作者：張麗麗，通訊方式：llychang93@gmail.com。

然而，近年來台灣學生參與各項國際性閱讀評量的表現並不理想，顯示國內閱讀教育推廣的成效仍有限。譬如：2006 年參與「國際學生能力評量計劃」(Programme for International Student Assessment，簡稱 PISA)，15 歲中學生閱讀素養在 57 個國家中排名 16，表現僅略高於「合作經濟發展組織」(Organisation for Economic Co-operation and Development，OECD) 的平均表現 (OECD, 2007) ；2009 年雖仍略高於 OECD 平均表現，但在 65 個參與國家及地區的排名卻降至 23，遙遙落後亞洲鄰國或地區 (OECD, 2010) 。在國小層級台灣學生的表現也不理想，2006 年「促進國際閱讀素養研究」(Progress in International Reading Literacy Study，PIRLS) 我國四年級學生的閱讀素養在 45 個國家及地區中排名 22，遠遠落後臨近的香港及新加坡 (柯華葳、詹益綾、張建妤、游婷雅，2008) 。

這些國際閱讀評量都將「閱讀素養」視爲面對與適應未來資訊社會所需具備的閱讀能力，是個人是否具備終身學習能力的重要指標，意即閱讀能力是一種解決問題及適應環境的能力。在這些閱讀評量中，閱讀理解被視爲一個複雜的認知活動。爲達理解目標，讀者需進行文字解碼、字面文意理解、整合前後文推論、整合文本與個人經驗之文意詮釋、對文章內容與結構進行批判與評鑑等各層次的理解活動。檢視台灣學生的表現，可以發現他們在較高層次的閱讀理解表現上相對較弱。譬如：PIRLS 2006 報告指出，臺灣學生在高層次「解釋理解歷程」的得分顯著低於較低層次「直接理解歷程」的得分，前者的通過率 49%也明顯低於後者的 73% (柯華葳等人，2008) ；PISA 2009 測驗結果也有相近的發現：台灣學生在「擷取與檢索」(access and retrieve) 、「統整與解釋」(integrate and interpret) 及「省思與評鑑」(reflect and evaluation) 三個層次達到六等級中五等級 (或以上) 的比例爲 9.9% (排名 18) 、6.4% (排名 26) 、4.9% (排名 33) (OECD, 2010) 。

由上述結果來看，提升我國學生的閱讀理解能力是一件刻不容緩的工作。除了在教學上應更強調高層次閱讀理解能力的培養外 (柯華葳等人，2008) ，閱讀理解評量也應能測量不同閱讀理解層次的表現，以做爲教師了解學生閱讀發展及改進閱讀教學的參考。國內目前雖有一些國語文成就測驗涵蓋閱讀理解的測量 (林寶貴、楊慧敏、許秀英，1995；邱上真、洪碧霞，1996，1997) ，但其評量重點多置於音韻轉錄或音韻處理、字形辨別、字義理解、語法理解、閱讀理解等各面向，並未側重不同理解層次的閱讀理解。有的雖爲閱讀理解測驗 (林建平，1994；胡永崇，1995) ，內容涵蓋字面理解、推論理解、摘要等重要閱讀理解成份，但因測驗編製目的在考驗特定實驗成效，故在信效度建立上並不周延。另有一些以診斷及篩選閱讀困難學童爲目的之閱讀理解測驗，譬如：「閱讀理解困難篩選測驗」(柯華葳，1999) 及「中文閱讀理解測驗」(林寶貴、錡寶香，2000) ，前者以短句或短文測部份處理及文本處理的能力、後者以短文測語言處理及閱讀理解次能力，但兩者皆非爲測一般學童之文本閱讀理解而設計。現有測驗中僅「國小學童閱讀理解測驗」(董宜俐，2003) 測一般學童不同層次之閱讀理解，惟此測驗只針對六年級學童，適用對象有限。有鑑於此，本研究主要目的在編製一份適用於國小高年級學童，測量不同閱讀理解層次之文本閱讀理解測驗。

一、閱讀理解歷程及測量

許多閱讀理論指出讀者在閱讀歷程中會形成不同層次的心理表徵。Kintsch (1988) 視閱讀理解爲一個複雜的認知歷程，讀者在閱讀時會形成微觀結構 (microstructure) 、鉅觀結構 (macrostructure) 及情境模式 (situation model) 等三種層次的理解表徵。微觀結構是指讀者由句子中抽取意義、產生命題，形成對篇章的初步理解，此層次屬於淺層的知識處理歷程。鉅觀結構是指讀者閱讀完文章後，整合所有文本的微觀結構，產生對篇章內容與文章主題的整體性理解，此層次屬於較深層的知識處理歷程。情境模式則是指讀者連結篇章內容與先備知識，形成對文本內容更高階的知識結構，這是更深層的知識處理歷程。Swaby (1989) 則將閱讀理解視爲一種技能，不同程度的閱讀技能可發展出四個閱讀層次：(1) 字義理解 (literal comprehension) 指讀者可從字句的語意，了解文章所欲明確表達的想法與主要概念；(2) 推論理解 (inferential comprehension) 指讀者能根據文章所描述訊息，加上自身經驗和直覺來推論文中隱藏的意涵；(3) 評鑑理解

（evaluative comprehension）指讀者能依據文章訊息產生自己的觀點；（4）批判理解（critical comprehension）則指讀者能分析文章中的寫作格式與內容。

整合 Kintsch（1988）和 Swaby（1989）的觀點，閱讀理解包含對文章句子表層意義的字面理解、對文章整體意義及隱含意義的推論理解、以及對文章內容及結構進行評鑑的批判理解。閱讀理解的評量應該含括這些閱讀理解層次，才能有效測量閱讀理解能力的全貌。

目前許多大規模閱讀評量皆結合理論及實徵研究，界定閱讀理解歷程並形成評量架構，譬如：PISA 界定「擷取與檢索」、「統整與解釋」和「省思與評鑑」三個層次（OECD, 2010）；PIRLS 界定「直接提取」（focus on and retrieve explicitly stated information）、「直接推論」（make straightforward inferences）、「詮釋、整合觀點與訊息」（interpret and integrate ideas and information）、和「檢驗、評估內容、語言和文章的元素」（examine and evaluate content, language, and textual elements）四個層次，前兩層次爲直接理解歷程、後兩層次爲解釋理解歷程（柯華葳等人，2008）。美國「國家教育發展評量」（National Assessment of Educational Progress，簡稱 NAEP）界定閱讀爲一動態認知歷程，包含「理解書寫文本」、「發展和解釋意義」及「依照文本類型、目的和情境適當地使用意義」（p.10），並以「尋找和回憶」（locate/recall）、「整合和解釋」（integrate/interpret）及「批判和評鑑」（critique/evaluate）三個認知標的層次做爲測驗編製架構（NAGB, 2008）。

NAEP、PIRLS 及 PISA 三者都是由認知歷程的觀點定義閱讀理解，其中 PIRLS 針對國小四年級學童、PISA 針對 15 歲中學生、NAEP 則針對四、八與十二年級學生。雖然中文在閱讀的識字歷程上有其獨特性（胡志偉、顏乃欣，1995；曾志朗，1991），但不論中外皆認爲閱讀的理解歷程包括字面理解、整合文意的組織、連結先備知識的精緻化等幾個次歷程，也因此 Kintsch（1988）模式廣爲國內閱讀研究採用，而各個跨國閱讀評量計畫皆採相近架構。本研究選取適用對象較廣的 NAEP 做爲編製閱讀理解測驗之依據。

二、NAEP 閱讀評量架構

NAEP 閱讀評量架構含文本類型及閱讀理解認知標的兩個面向（表 1）。文本類型分文學及訊息兩類，前者含小說、非小說文學、詩等三類，後者含說明文、議論及論說文、文件及程序等三類。在認知標的層次方面，「尋找和回憶」評量文本中小部份（如：一個句子、一個或相鄰少數段落）的明述訊息，屬閱讀理解最基本的技巧；「整合和解釋」指讀者能以完整且合乎邏輯的方式進行文本中大部份、整體，甚至跨文本的整合及解釋，此時讀者已超越文本所提供之間斷訊息、觀點、細節、話題等，在一個較抽象的層次提問、形成意象或連結，至此階段，讀者已能連結文本訊息及先前所學或經驗、爲特定目的或需要而閱讀、應用閱讀所得新知於真實生活中；「批判和評鑑」指讀者能跳脫文本，從多元觀點綜合其他資料或經驗，客觀的批判文本。本文參考上述評量架構選取文本及編寫試題。

NAEP 對不同文本類型、認知標的及題型的比重分配，隨年級上升而增加訊息文本、高層次認知歷程、開放題型的比例。文學文本及訊息文本之試題在四年級各佔 50%，在八年級各佔 45%及 55%；「尋找和回憶」、「整合和解釋」及「批判和評鑑」在四年級的比重爲 30%、50%、20%，在八年級爲 20%、50%、30%；選擇題及建構型試題在四年級各佔 50%（其中 10%爲延伸性建構反應題）、在八年級各佔 40%與 60%（其中 15%爲延伸性建構反應題）。本研究受試對象爲五、六年級學生，故除參考上述比例外，亦根據我國教學現場評量使用情形及受試者預試反應，安排文本類型、認知標的及題型的比重（見「研究方法」中測驗內容的說明）。

表 1 NAEP 閱讀評量架構

文本類型	閱讀理解認知標的層次		
	尋找和回憶	整合和解釋	批判和評鑑
文學、訊息文本共同特質	確認文本內明確訊息、做文本內及跨文本之簡單推論： ・定義 ・事實 ・支持細節	進行文本內及跨文本之複雜推論： ・描述問題和解決方式、因果關係 ・觀點、問題、情境的比較或連結 ・論述中未明述之假定 ・描述作者如何使用文學手法和文本特徵	以批判方式： ・判斷作者的手法與技巧 ・評鑑作者在文本內或跨文本的觀點及看法 ・對文本持不同觀點
文學文本特質	確認文本內明確訊息、做文本內及跨文本之簡單推論： ・角色特質 ・事件和動作順序 ・場景 確認比喻性語言	進行文本內及跨文本之複雜推論： ・推論情緒或語調 ・整合觀點判斷主題 ・確認或解釋角色的動機和決定 ・檢視主題、場景、角色的關係 解釋節奏、韻、形式如何建構詩的意義	以批判方式： ・評鑑文學手法在傳達意義上所扮演的角色 ・判斷文學手法提升文學作品的程度 ・評鑑角色的動機及決定 ・分析作者使用的觀點
訊息文本特質	確認文本內明確訊息、做文本內及跨文本之簡單推論： ・主題句或主要概念 ・作者目的 ・因果關係 在文本或圖表中找出特定訊息	進行文本內及跨文本之複雜推論： ・摘要主要概念 ・做出結論和提出支持訊息 ・找到支持立論的證據 ・區別事實與意見 ・判斷文本內及跨文本訊息的重要性	以批判方式： ・分析訊息的呈現 ・評鑑作者選擇語言來影響讀者的方式 ・評鑑作者用來支持立論之證據的品質與強度 ・決定文本內及文本間反駁論點之品質 ・評鑑立論之連貫性、邏輯、可信度

摘自"Reading Assessment and Item Specification for the 2009 National Assessment of Educational Progress," NAGB, 2008, p. 46.

三、Rasch 客觀測量模式編製「國小高年級閱讀理解測驗」

目前國內閱讀理解測驗普遍以古典測驗理論（CTT）編製測驗（如：根據試題難度、鑑別度等指標篩選試題、以原始分數建立信效度憑證等），惟根據 CTT 建構之工具得到的原始分數並不符合「測量」原則，不爲等距量尺（即選定的「單位尺度」在潛在量尺上所表示的量不是恆定的），因此在解釋與比較個體之潛在特質上有其限制（王文中，1996；Andrich, 1988; van der Linden, 1994）。Rasch「明確客觀」測量（specific objectivity, Rasch, 1960/1980）之測量工具與被測量的客體相互獨立，意即：在控制受試者能力的情形下，試題難度的差異只與試題本身的難度有關，不受受試者能力、試題鑑別度或其它因素的影響；在控制試題難度的情形下，受試者能力的差異也只與受試者本身的能力有關，與試題難度或其它因素無關。此時，試題參數(δ)與能力參數(β)具可加性的關係，可置於單一向度的連續量尺上直接比較，以二元計分模式爲例：$\ln\left(p_{ni}/(1-p_{ni})\right)=\beta_n-\delta_i$。惟須注意達到客觀測量的前提是數據必須契合 Rasch 單向度模式。要具備明確客觀特性，測量必須符合兩個條件：對任何能力水平的受試者，能力越高答對試題的機率越高（試題特徵曲線 ICC 呈單調遞增）、受試者在簡單試題的答對機率要高過困難試題的答對機率（即試題難度排序不變）；所有 IRT 模式均符合前項條件，但僅 Rasch 模式符合後項條件（王文中，1996）。簡言之，不同模式的立論及目的不同，Rasch 模式爲測量模式，強調數據是否契合

理論的單向度模式；其它 IRT 模式則爲統計模式，強調找出最契合數據的模式，因此，在加入鑑別度及其他參數的情形下，試題對不同能力受試者而言，其簡單或困難的排序是不同的。

CTT 及不同 IRT 模式的立論及目的不同，各有利弊，基於「客觀測量」特性，Rasch 模式在建置閱讀理解測驗上有一些優點。（1）可檢核閱讀理解構念的向度性（dimensionality），即檢核不同閱讀理解成份或認知層次是否隸屬單一構念向度；由於閱讀理解表現受文本類型及文章複雜度的影響甚大，故根據 CTT 建置之測驗甚少或難以透過因素分析的方式檢核閱讀理解的構念結構（通常會呈現「文本」或「難度」因素，而非閱讀認知向度）。（2）當閱讀理解構念符合單向度時，我們可以探討閱讀理解認知標的之層次或階層性；目前 CTT 取向的方法大多根據不同層次試題的平均難度間接推論閱讀理解的認知層次是否符合理論上的預期（屬於「間接」推論的原因在於難度與構念並無直接關係）。（3）因受試者能力可直接與試題難度相比較，故可推論受試者回答個別試題及達到特定認知層次的機率，並利用此訊息診斷學生閱讀理解的認知層次；CTT 依據之測驗則因爲整體測驗取向，故無法根據受試者的測驗總分推論其在個別試題或特定認知層次上的表現。（4）透過「能力與難度對應圖」（亦稱 Wright Map）選擇廣度（涵蓋不同能力）與精確性（與受試者能力相近）兼顧的試題；CTT 取向的方法因主要依據試題間相關（如：試題與總分相關）來選題，故同質性高的試題往往因能提高精確性（信度）而成爲較佳的選擇，然此舉通常會降低構念的廣度及代表性（Engelhard, 1993; Singh, 2004）。（5）閱讀理解測驗通常包含混合題型或不同等級計分之試題來評量不同認知層次之閱讀理解，Rasch 模式因試題難度及受試者能力估計值不受計分等級數的影響，故可輕易解決此問題，但 CTT 取向的方法則較難處理此問題（如：不易決定不同試題之比重）。（6）當測驗包含建構反應試題時，可透過等級閾值（thresholds）檢驗計分等級的階層性（不同等級是否反映不同程度的構念特質）及適切性（等級間距是否恰當，不致於過窄而無法區辨受試者差異，也不致於過寬而降低了測量的精確性），並可根據結果適度調整等級數，CTT 則未提供考驗這些特質的機制。其它優點尚包括：Rasch 模式在建構測驗可同時將「差異試題功能」（differential item functioning，簡稱 DIF）及評分者信度等一併納入考驗。（有關 Rasch vs. CTT 模式探討心理計量特性的討論可參見張麗麗、羅素貞，2011）

綜合上述，本研究目的在參考 NAEP 閱讀理解評量架構，以 Rasch 模式編製一份適合國小五、六年級一般學童的文本閱讀理解測驗、建立各項信度與效度的憑證、並提出測驗分數的可能應用方式。

方法

一、研究對象

研究對象爲高屏地區國小五、六年級學童，含預試、正式施測及效標樣本。

（一）預試樣本

預試分爲三階段。第一階段含五次小預試，對象爲高雄市一所國小五、六年級不同語文能力之學生 12 位，目的在確定文本及試題的適切性。第二階段以立意取樣，根據學校規模從高屏地區選取五所國小（大、中、小型各 2、2、1 校），每校五、六年級各兩班，共 20 班 458 位學生，目的在以 Rasch 模式檢視試題的計量特性。之後根據第二階段預試結果，修正文本及試題，並進行第三階段預試以確定各項估計值的穩定性，對象爲參與第二階段之一所大型學校（五、六年級各一班）及一所中型學校（五、六年級各二班），共六班 176 位學生。

（二）正式施測樣本

本研究根據 98 學年度學校資料，以高屏地區五、六年級學童爲母群，依地區、學校規模及年級比例，分層隨機抽取 15 校 40 班 1114 位學生（若欲抽取人數不足一班時，仍取一班），剔除無效卷得有效樣本 1052 份。其中，高雄市、高雄縣、屏東縣人數分別爲 451（42.9%）、360（34.2%）、241（22.9%），大、中、小型學校人數爲 399（37.9%）、533（50.7%）、120（11.4%），五、六年級人數爲 526（50.0%）及 526（50.0%），男、女人數爲 534（50.8%）、518（49.2%）。

（三）效標樣本

研究者自正式樣本中選取一所中型學校的五、六年級各三班，計 165 位學生，蒐集外在效標（中文閱讀理解測驗、國語文及數學學期成績）並建立評分者信度。

二、「國小高年級閱讀理解測驗」

（一）編製程序

本研究參考表 1「NAEP 閱讀評量架構」建置測驗。編製之初，先從多種來源搜尋適當文本，初步選定及編修後，由本文第一作者、兩位高年級教師及一位測驗專家評估內容的適切性及用字難易度；其後，選取五、六年級之高、中、低語文能力學童 12 位進行個別訪談，了解文本的適切性。最後，選出合適的五篇文本，爲考量高年級已由「學習如何閱讀」進入「透過閱讀學習」階段，故選定三篇訊息文本及兩篇文學文本。所有文本皆透過「中文文章適讀性線上分析系統」（高師大工業科技教育系，2009）進行文本長度及用字難度的可讀性分析。

文本確定後，建立測驗明細表、編寫試題並完成初稿。其後進行第一階段小預試，檢核文本及試題用字、題意及選項，並由四位修習過閱讀認知及測驗編製課程，且參與過一個學期閱讀理解測驗編製工作坊之研究生兼國小教師檢視測驗內容以及其與明細表之配合度。之後，根據建議修改，完成預試版，共五個文本 47 題。

接著進行第二階段預試，以 Rasch 模式分析測驗向度、試題難度、模式契合度、誘項適切性、建構反應題之計分等級適切性、信度等，同時對學生建構反應題之作答進行質化分析，以建立計分規準。之後，根據結果小幅修正無法引發預期反應之文本內容並刪修試題，保留 41 題。最後再次由預試之四位審查者、閱讀專家及測驗專家各一位檢核測驗品質。爲確保修改後測驗計量特性的穩定性，進行第三階段預試及 Rasch 分析，結果大致穩定，得正式施測版 41 題。

正式施測後發現有兩題與 Rasch 模式的契合度不理想，故決定刪除，最後保留 39 題，並依據此建立信效度憑證、計算轉換量尺分數與建立常模。

（二）測驗內容

測驗含五篇文本 39 題，其中 28 題爲對/錯計分的四選一選擇題、11 題爲建構反應實作題（performance assessment，PA）。爲方便施測，五篇文本裝訂爲甲、乙兩題本。表 2 爲根據文本、題本及認知層次分類的測驗明細表，表 3 爲試題認知層次、題型及認知標的對照表。

在文本及題本方面，二篇爲文學文本，含傳記文學類「瑞秋·卡森的故事」（A 文本，8 題，其中 3 題爲 PA）、生活故事類「記憶的項鍊」（C 文本，8 題，其中 2 題爲 PA），三篇爲訊息文本，含說明文 「不再旅行的鮭魚－櫻花鉤吻鮭」（B 文本，9 題，其中 3 題爲 PA）及「地球發燒了」（E 文本，6 題，其中 1 題 PA）、說明文/故事混合類「巧克力最早是『苦水』飲料？」（D 文本，8 題，其中 2 題爲 PA）。甲題本含 A、B 文本（17 題，6 題爲 PA）、乙題本含 C、D 及 E 文本（22 題，5 題爲 PA）；爲方便學童閱讀及作答，文本與試題分開裝訂。

在認知層次方面，尋找和回憶、整合和解釋、批判和評鑑等三層次所佔比重爲 36%、51%、13%，與 NAEP 國小階段三層次 3：5：2 的比重略有出入，主要原因有二：一是本測驗訊息文本較多，受文本特性影響，層次一題目比重較高；二是考量國內目前正式評量仍少建構反應題，而層次三又以此類題型爲主，故在參考學生預試之作答反應（作答意願低、空白或敘述不完整等）後，調降層次三的比重。

在題型方面，選擇題及建構反應題的比重爲 72%（28 題）及 28%（11 題），與 NAEP 四年級兩種題型各佔 50%的比重有出入，其原因如前所述。

表 2 測驗明細表

文本及題本	認知層次			總計題數
	尋找和回憶	整合和解釋	批判和評鑑	
文學文本 (A, C 文本)	5	9	2	16 (5)
訊息文本 (B, D, E 文本)	9	11	3	23 (6)
題本甲 (A, B 文本)	4	11 (4)	2 (2)	17 (6)
題本乙 (C, D, E 文本)	10	9 (2)	3 (3)	22 (5)
總計題數	14	20 (6)	5 (5)	39 (11)
%	36%	51%	13%	

註：文本 A「瑞秋·卡森的故事」、B「不再旅行的鮭魚－櫻花鉤吻鮭」、C「記憶的項鍊」、D「巧克力最早是『苦水』飲料？」、E「地球發燒了」。題數右側（）內爲建構反應題數。

表 3 試題認知層次、題型及認知標的

題號	認知層次	題型	認知標的	題號	認知層次	題型	認知標的
A1	2	MC	整合觀點判斷主題 (跨段落)	C3	2	MC	解釋主角動機 (連結經驗)
A2	1	MC	簡單推論 (跨段落)	C4	1	MC	事件/動作順序 (跨段落)
A3	2	MC	連結情境與經驗 (連結經驗)	C5	1	MC	角色特質 (跨段落)
A4	1	MC	確認比喻性言語 (段落內)	C6	1	MC	找出特定訊息(跨段落重複訊息)
A5	2	MC	檢查主題/場景/角色關係(跨段落)	C7	2	PA	推論情緒(跨段落+連結經驗)
A6	2	PA	推論角色情緒 (連結經驗)	C8	3	PA	判斷作者手法/技巧效果(跨段落)
A7	2	PA	整合概念決定主題 (跨段落)	D1	2	MC	根據訊息推論 (跨段落)
A8	3	PA	判斷作者動機 (跨段落)	D2	1	MC	找出特定訊息 (跨段落)
B1	1	MC	找出特定訊息 (段落內，但其他段落含相似訊息)	D3	1	MC	找出支持細節 (跨段落)
				D4	1	MC	找出支持細節 (跨段落)
B2	1	MC	找出特定訊息 (段落內)	D5	1	MC	特定訊息 (跨段落)
B3	2	MC	連結/比較訊息 (跨段落)	D6	2	MC	摘要主要概念 (段落內)
B4	2	MC	連結/比較訊息 (跨段落)	D7	2	PA	摘要整篇概念發展 (跨段落)
B5	2	MC	推論觀點 (跨段落)	D8	3	PA	判斷作者選擇語言目的(跨段落)
B6	2	MC	連結情境與經驗 (連結經驗)	E1	1	MC	根據重複訊息推論主題(跨段落)
B7	2	PA	摘要主要概念 (段落內)	E2	1	MC	圖表特定訊息 (單一圖表內)
B8	2	PA	統整/比較訊息 (跨段落)	E3	2	MC	圖表整合與推論 (跨圖表)
B9	3	PA	評斷作者使用語言 (跨段落+連結知識與經驗)	E4	1	MC	根據重複訊息推論主要概念(跨段落)
C1	2	MC	整合概念決定主題 (跨段落)	E5	2	MC	區辨事實與意見 (跨段落)
C2	2	MC	解釋主角動機 (跨段落)	E6	3	PA	分析訊息及圖表呈現 (跨段落)

註：認知層次 1 = 尋找和回憶、2 = 整合和解釋、3 = 批判和評鑑。題型 MC = 選擇題、PA = 建構反應實作題。

（三）計分

所有選擇題皆以對/錯計分，建構反應題除 B7 爲三等級外，皆爲二等級（原有 4 題爲三等級，後因 3 題在「等級適切性」之閾值分析不理想，而改爲二等級計分，見「研究結果」部份）。測驗原始總分介於 0~40 分，量尺分數介於 0~100 分（平均數 50、標準差 10.87）（見研究結果「測驗分數的轉換與應用」）。

建構反應題之計分規準如表 4。研究者參考實作評量計分規準之建構步驟，先以質化方式分析學童的作答反應，再根據試題訊息量多寡，擬出兩等級（0~1）及三等級（0~2）兩種計分規準。因唯一採三等級計分之 B7 含兩個單獨以 0~1 計分的子題（單獨計分後加總），故本研究僅採 0~1 計分規準。研究者也針對各題找出不同作答型態的範例，經與測驗專家討論及試評後，做爲評分範本。

表 4　建構反應試題計分規準

等級		規準內容
1	理解	‧作答內容清楚正確，可以證明對文本和題目有理解。 ‧能從文本中找出有關且正確的訊息來支持自己的答案。
0	極少或未理解	‧作答內容模糊、不完整、矛盾或錯誤，無法證明對文本或題目有理解。 ‧以文本中不適合、不正確或無關的訊息來支持自己的答案。 ‧僅以個人無關的觀點與想法來回答問題，未能從文本中找出有關且正確的訊息來支持自己的回答。 ‧僅複述題目，未提出個人看法。 ‧空白

爲建立建構反應題的評分者信度，效標樣本的試卷由本文第一作者及另一位具 15 年國小教學經驗，且具閱讀及測驗背景之研究生共同計分；評分前先進行評分者訓練，討論計分規準及評分範本、依據計分規準反複試評，直至兩人計分一致，才分開獨立計分。校標樣本外的所有建構反應題，皆由本文第一作者評分，因計分規準及評分範本明確，且根據效標樣本得到之評分者信度高（見「研究結果」），故評分結果應具一定程度的可信度。

（四）施測程序

測驗以標準化方式施測，由研究者提供施測指導語及注意事項予各班導師，由其在原班教室內利用早自修或上課時段施測。爲避免受疲憊因素的影響，兩份題本於一週內不同的兩天施測。另外，考量測驗題本施測順序對學生表現的可能影響，一半班級先測甲題本，再測乙題本，另一半則以相反順序施測。

三、其它測量工具

（一）中文閱讀理解測驗

「中文閱讀理解測驗」由林寶貴與錡寶香（2000）編製，適用國小二至六年級學童，主要目的爲篩選閱讀理解困難之學童或推論身心障礙學生閱讀理解能力。測驗包含故事類記敘文與說明文各六篇，共 100 題對/錯計分的選擇題，測驗主要評量語言處理能力（音韻處理、語法、語意）及閱讀理解能力（理解文章基本事實、摘要重點大意、推論、比較分析）。在信度方面，全測驗兩週重測信度 .89、庫李信度 .90、折半信度 .95；不同年級內部一致性信度係數介於 .88 ~ .96；在效度方面，閱讀理解次能力之間的相關介於 .79 ~ .98，全測驗與次能力之相關介於 .85 ~ .96，全測驗與「中華國語文能力測驗」全測驗之相關爲 .80。

（二）國語文與數學學期成績

研究者取得效標樣本學童 97 學年度下學期的國語文與數學學期成績，並以班級爲單位，轉換爲標準 *z* 分數。

四、資料分析

（一）Rasch 模式

由於測驗包含二元及多元計分試題，故採 Rasch 部份給分模式（partial credit model）（Master, 1982）：

$$\pi_{xni} = \frac{\exp\sum_{j=0}^{x}\left(\beta_n - \delta_{ij}\right)}{\sum_{k=0}^{m_i}\exp\sum_{j=0}^{k}\left(\beta_n - \delta_{ij}\right)}$$

π_{xni} 爲受試者 n 在試題 i 達到等級 x 的機率，β_n 爲受試者能力、δ_{ij} 爲試題 i 第 j 個閾值的難度（閾值又稱 step 難度，爲相臨兩等級機率相等之值）。$x=0,1,...,m_i$（等級數爲 $m+1$）、$k=1,2,...,m_i$（閾值數相當於等級數減 1），$\sum_{j=0}^{0}\left(\beta_n-\delta_{ij}\right)\equiv 0$ 。

本研究主要採 RUMM2020 軟體（Andrich, Sheridan, & Luo, 2007），此軟體以 PAIR 估計試題參數、JMLE 估計能力參數；由於 RUMM 僅提供統計考驗依據之試題契合度指標，爲避免統計考驗受大樣本影響而過度敏感，研究者亦採 WINSTEPS 軟體（Linacre, 2005）所提供較不受樣本數影響之 infit MNSQ 及 outfit MNSQ 契合度指標（此軟體採 JMLE 估計參數）。

由於閱讀理解測驗由題組試題組成，同一文本內的試題可能存在局部依賴（local dependence）現象，而造成參數估計的偏誤（Smith, 2005; Yen, 1993），故本研究以殘差相關（相當於 Yen 所提的 Q3 指標）檢視此假定，若殘差相關高則表示在 Rasch 因素被解釋後，試題反應間仍有相關存在，表示局部獨立性被違背（Wright, 1996），此時，試題不宜以獨立試題分析。本研究殘差相關分析的結果顯示在 39 題的 780 對殘差相關中，97.4%的相關小於 ±.10 ，2.6%的相關大於 .10（其中僅一對高過 .20），顯示本測驗試題雖有共同文本，但題目間仍具相當程度的獨立性，故以個別試題分析應屬適切。

（二）效度分析

以下從測驗內容、內在結構、與外在變項關係等三方面建立效度憑證。

1. 測驗內容

研究者透過建立測驗明細表、專家檢核、預試分析等方式確認與提升測驗內容之關聯性與代表性。

2. 內在結構

本研究檢核試題等級適切性、構念向度性、年級及性別 DIF、試題難度適切性及閱讀理解認知層次之階層性等。

在等級適切性方面，檢核多元計分試題的閾值是否具階層性、各選項次數是否大於 10（過少影響參數估計之穩定性）、閾值間距是否介於 1.0 ~ 5.0 logits 的適切間距（Linacre, 2004）。

在構念之單向度檢核方面，以 outfit 及 infit MNSQ 檢視個別試題契合度。MNSQ 值介於 0 至無限大，期望值爲 1，大於 1 表示試題不契合單向度模式、小於 1 表示數據與模式過度契合，二元及多元計分試題的合理範圍分別爲 0.70 ~ 1.30、0.60 ~ 1.40（Wright & Linacre, 1994）。另外採殘差主成份分析的第一個特徵值做爲整體契合度指標，當數據契合 Rasch 模式，殘差應隨機且獨立，若 Rasch 因素被解釋後，殘差仍可抽取出因素，就表示所測構念可能含多個向度。Smith 與 Miao（1994）的模擬數據研究顯示當單向度假定契合時，第一個主成份的特徵值低於 1.4；本研究根據 Rasch 理論模式產生模擬數據，再比較從實徵及模擬數據得到之特徵值的差異，若兩者差異不大，則表示數據未違背單向度假定。

在年級及性別 DIF 檢定方面，任何測驗之構念均不應群體不同而有差異，意即具相同能力但來自不同次群體之受試者，其答題機率應無不同（Holland & Thayer, 1988）。本研究以殘差雙因子變異數分析及試題期望 ICC 檢核 DIF。殘差雙因子變異數分析中「能力組別」及「DIF 變項」爲主要因子，當「DIF 變項」達顯著但「能力組別×DIF 變項」交互作用未達顯著時，DIF 爲齊一性（uniform DIF）；「能力組別×DIF 變項」達顯著時，DIF 爲非齊一性（non-uniform DIF）[第一類錯誤控制率 $\alpha = 0.05/(39\times2) = .00064$]（Andrich, Sheridan, & Luo, 2007）。若試題呈現齊一性 DIF，在單向度符合的情形下，研究者將試題拆成年級特定或性別特定試題，譬如：某題呈現性別 DIF，則拆成男生及女生兩個試題，此作法在 PAIR 估計法下，相當於以等化連結試題（Andrich & Hagquist, 2004）。

在試題難度及適切性方面，檢視不同文本及各類型試題之難度，並以 Wright Map 檢視試題難度分配的廣度及其與受試者能力對應的情形。

在閱讀理解認知層次之階層性方面，以 Wright Map 及單因子變異數分析檢核不同閱讀理解認知層次的難度是否符合理論所預期之階層。

3. 與外在變項關係

本研究以聚斂/區辨相關、群體差異來檢視閱讀理解測驗與外在變項之關係：以學童 Rasch 能力估計值與「中文閱讀理解測驗」（林寶貴、錡寶香，2000）、國語文及數學學期成績（*z* 分數）之相關，建立聚斂與區辨相關；以獨立樣本 *t* 檢定考驗五、六年級及男、女學童之閱讀能力估計值之差異；以單因子變異數分析考驗不同國語文能力學童閱讀能力估計值之差異。

（三）信度分析

本研究建立受試者分隔信度及評分者信度。受試者分隔信度係數爲受試者變異與真分數變異（即 Rasch 單向度模式可解釋之變異）之比值，係數愈高表示試題愈能穩定區隔受試者差異，受試者的排序愈穩定（Smith, 2001）。評分者信度則以 Rasch 多面相模式（many-facet model）估計兩位評分者的嚴苛度。

結果

一、效度分析

（一）測驗內容之憑證

研究者透過建立明確 NAEP 依據測驗明細表，由閱讀領域及測驗專家以邏輯分析方式，佐以各次預試分析結果，不斷檢核測驗內容與明細表之配合度，並進行修正，故測驗內容與閱讀理解應具關聯性，且爲閱讀理解構念之代表性樣本。

（二）內在結構之憑證

1. 等級適切性

本測驗 11 題建構反應題中有四題（B7, B8, B9, C8）採三等級計分，閾值分析結果顯示僅 B7 閾值理想，故維持三等級計分，其餘皆改以二等級（0~1）計分（B8 閾值間距過窄不及 1.0 logit、C8 的閾值次序顛倒、B9 間距理想，但最高等級之次數不及 10，可能影響參數估計之穩定性）。

2. 單向度檢核

在個別試題契合度方面，infit MNSQ 介於 0.81 ~ 1.13（平均值 1.00、標準差 0.07）、outfit MNSQ 介於 0.75 ~ 1.20（平均值 0.99、標準差 0.12），符合二元（0.7 ~ 1.3）及多元（0.6 ~ 1.4）計分試題的合理範圍，顯示試題契合單向度模式，且無過度契合現象。在整體契合度方面，殘差主成份分析的結果顯示最大特徵值爲 1.8，僅略高於根據單向度模擬數據得到的最大特徵值 1.4。綜合個別試題和整體測驗契合度指標，本測驗契合單向度模式，閱讀理解可由單一構念解釋。

3. 年級及性別 DIF 檢定

殘差雙因子變異數分析結果顯示部份試題出現年級及性別「齊一性」DIF，爲避免估計值受其它試題影響，研究者從最嚴重的 DIF 試題開始逐題處理，將同一題目拆成年級特定（或性別特定）的兩個試題，再重新進行 DIF 分析，循環此程序，直至所有試題皆未出現 DIF 才停止。最後，共確認二題性別 DIF 試題（C6、C7），皆對女生有利；三題年級 DIF 試題（B1、B8、B9），皆對六年級有利。

檢視 DIF 試題之內容，發現兩題性別 DIF 試題皆來自「記憶的項鍊」一文，此兩題均要求學生以較細膩的心思作答，譬如：C7 請學生判斷爲何當故事中女孩因懷念過世母親而對新媽媽在言語上不客氣時，她的父親並未生氣的理由，因文章並未對女孩父親做太多的敘述，故學生必須連結自身感受作答，由於女生情感表達發展較男生早，故這兩題皆對女生較有利。圖 1 顯示女生之實際 ICC（不平滑曲線）高過男生之實際 ICC，表示對具相同閱讀理解能力的學童言，女生答對機率高過 Rasch 模式之預期（平滑曲線），而男生則低過模式預期。三題年級 DIF 試題皆來自「不再旅行的鮭魚－櫻花鉤吻鮭」文本，這三題均要求學生處理數個段落中諸多訊息後再作答，所測內容的認知負荷量對五年級學童可能過重，造成對六年級較有利的現象。

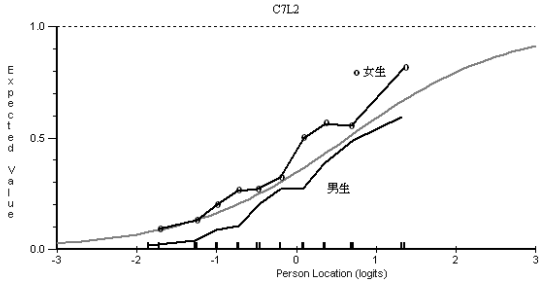


圖 1　性別 DIF 試題範例

確定 DIF 試題後，研究者將二題性別 DIF 試題拆成四題（男生及女生各二題）、三題年級 DIF 試題拆成六題（五年級及六年級各三題）。注意總試題數雖變爲 44 題（原 39 題加上拆題多出的 5 題），但每位受試者的能力估計值仍然是依據其所屬次群體之 39 題而定。

4. 試題難度及適切性

表 5 爲試題難度及受試者能力（個別試題難度參見表 10）、表 6 爲不同分類試題之平均難度與差異考驗、圖 2 爲受試者能力與試題難度對應之 Wright Map。圖 2 最左邊「位置」標示量尺 logit 值、圓點標示受試者能力分配、右邊爲各文本試題之難度分配；量尺愈靠上方表示受試者能力愈高或試題難度愈高、愈靠下方表示受試者能力愈低或試題難度愈低；當受試者能力與試題難度相同時，受試者有 50%的答對機會，當受試者能力高過試題難度時，受試者有 50%以上的答對機會，反之，則不及 50%（試題編碼見圖 2 註解）。

首先，看整體測驗，試題難度介於-2.94 ~ 1.87、M = 0.00，學童能力介於-2.42 ~ 3.00、M = -0.24，試題平均難度略高於學童平均能力；Wright Map 顯示試題難度與學童能力的對應佳，不同能力學童皆有相對應難度之試題。

表 5　試題難度及受試者能力

	試題難度	受試者能力
平均值	0.00	-0.24
標準差	0.95	0.92
最小值~最大值	-2.94 ~ 1.87	-2.42 ~ 3.00

表 6　不同分類試題之平均難度與差異考驗

試題分類	題數	平均難度	標準差	最小值	最大值	難度差異考驗
文本類型						
1. 文學	18	-0.20	1.09	-2.94	1.08	t（42）= 1.17
2. 訊息	26	0.14	0.84	-1.56	1.87	p = .25
文本						
A. 瑞秋.卡森的故事	8	0.22	0.73	-1.46	0.82	
B. 不再旅行的鮭魚-櫻花鉤吻鮭	12	0.21	0.79	-1.56	1.54	
C. 記憶的項鍊	10	-0.38	1.32	-2.94	1.08	
D. 巧克力最早是『苦水』飲料？	8	0.32	1.04	-1.39	1.87	F（4，39）= 0.85
E. 地球發燒了	6	-0.24	0.65	-0.91	0.97	p = .50
試題類型						
1. 選擇題	30	-0.41	0.82	-2.94	0.63	t（42）= 5.40
2. 建構反應題	14	0.88	0.53	0.05	1.87	p = .00
認知層次						F（2，41）= 31.65
1. 尋找和記憶 (L1)	16	-0.89	0.84	-2.94	0.17	p = .00
2. 整合和解釋 (L2)	22	0.32	0.44	-0.32	1.59	Scheffé事後比較：
3. 批評和評鑑 (L3)	6	1.20	0.41	0.82	1.87	L3 > L2 > L1

位置	受試者	試題 [uncentralised thresholds]				
3.0		瑞秋(A) (文學)	鮭魚(B) (訊息)	項鍊(C) (文學)	巧克力(D) (訊息)	地球發燒(E) (訊息)
2.0	o					
	o				D8L3P	
	o					
	oooo		B9L3P(G5)		D7L2P	
	oooooo					
1.0	ooooooo			C8L3P,C7L2P(B)		
	oooooooo	A8L3P	B7L2P.2,B9L3P(G6)			E6L3P
	oooooooo		B6L2,B8L2P(G5)			
	ooooooooooooo	A5L2P,A7L2P				
	oooooooooooooo	A6L2P	B3L2	C2L2,C7L2P(G)	D1L2,D6L2	
0.0	ooooooooooooooo	A1L2	B8L2P(G6),B4L2,B1L1(G5)	C3L2,C1L2	D5L1,D4L1	
	oooooooooooooo		B7L2P.1			E5L2
	ooooooooooooo	A3L2	B5L2		D3L1	E3L2
	ooooooooooooooo	A2L1	B1L1(G6)			E2L1
	oooooooooooooo					E4L1
−1.0	ooooooooooooo			C5L1,C4L1		E1L1
	oooooooooooooooo					
	oooooooo				D2L1	
	ooooooo	A4L1	B2L1			
	ooooo					
−2.0	ooo					
	ooo			C6L1(B)		
	oo					
−3.0				C6L1(G)		

o=5 Persons (人數低於5人者未列出)

圖 2　受試者能力與試題難度 Wright Map

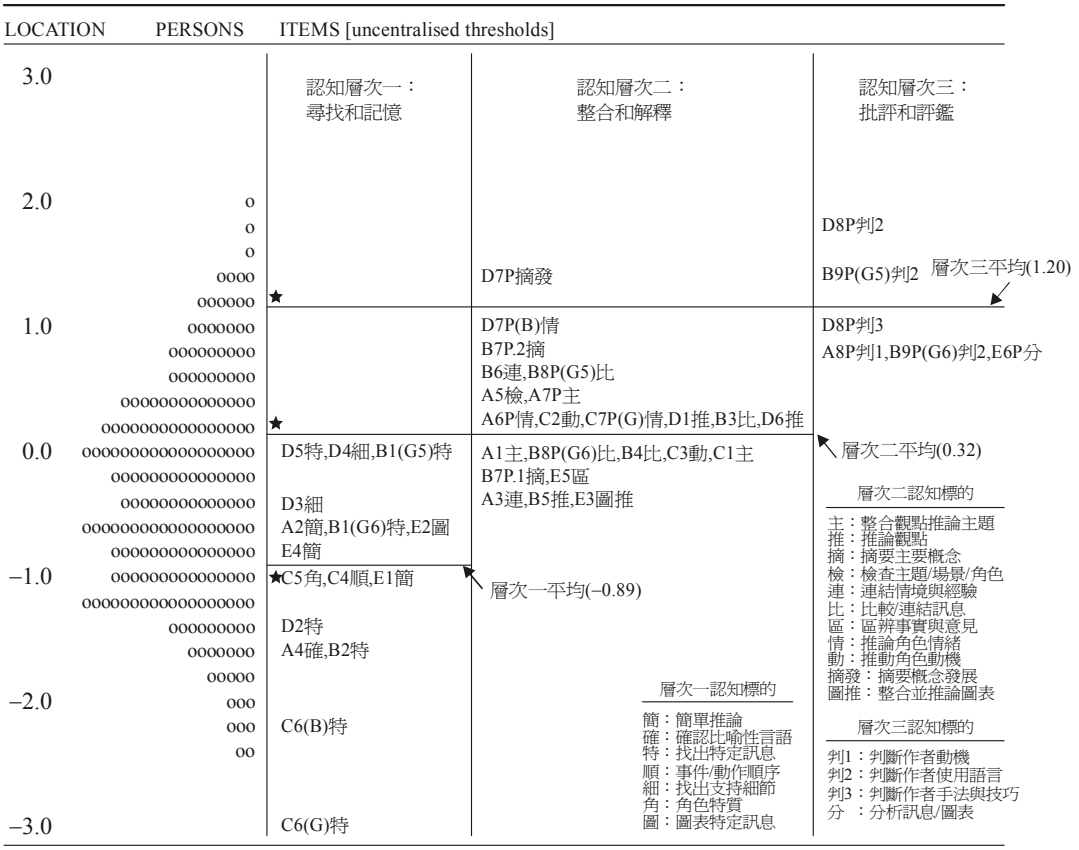
註：題目編碼 1-2 碼爲題本及題本內編號、3-4 碼爲認知層次，題號後註明 P 者爲建構反應題，題號句點後數字標示第 k 個閾值，(　)內標示 DIF 試題拆題後所隸屬的次群體（如：G5 爲五年級試題、G6 爲六年級試題、B 爲男生題、G 爲女生題）。舉例說明：B9L3P(G5)爲 B 題本第 9 題、測認知層次 3、建構反應題(P)、屬 5 年級試題；B7L2P.2 爲 B 題本第 7 題、測認知層次 2、建構反應題第二個閾值(P.2)。

其次，從文本及題型看試題難度分佈（表 6 及圖 2）。文學文本試題難度分佈（-2.94 ~ 1.08）較訊息文本廣（-1.56 ~ 1.87），平均難度（M = -0.20）略低於訊息文本（M = 0.14），但差異未達顯著（p = .25）。五個文本中「記憶的項鍊」難度最低（M = -0.38、介於-2.94 ~ 1.08）、「巧克力最早是『苦水』飲料？」難度最高（M = 0.32、介於-1.39 ~ 1.87），此兩文本試題的分佈較其他文本廣泛，比較適合測能力分配廣的群體；其他文本試題的全距較小，平均難度接近整體測驗的平均難度（0.0），比較適合測中等能力的學童；五個文本的平均難度未達顯著差異（p = .50）。建構反應題（圖 2 標示 P）的難度（介於 0.05 ~ 1.87、M = 0.88）顯著高於選擇題的難度（介於-2.94 ~ 0.63、M = -0.41）（p = .00），前者適合測中、高能力學童，後者適合測中、低能力學童；能力達建構反應題平均難度（0.88）的學童，答題勝算率是能力達選擇題平均難度（-0.41）學童的 3.63 倍（$e^{(0.88-(-0.41))} = 3.63$）。

整體言，試題難度的分佈廣且與受試學童的能力相對應，不同文本的分佈情形略有不同，但不論是訊息或文學文本，其平均難度均無顯著差異；建構反應試題難度高過選擇題，兩者分別適合測不同能力的學童。

5. 閱讀理解認知層次之階層性

閱讀理解層次之階層性可從表 6、圖 2 及圖 3 來檢視。



o=5 Persons (人數低於5人者未列出)

圖 3　受試者能力與不同認知層次試題難度 Wright Map

註：圖中右下方縮寫標示試題之認知標的。題號編碼：1-2 碼爲題本及題本內編號、P 爲建構反應題、句點後數字標示第 k 個閾值、(　)內標示 DIF 試題拆題後所隸屬的次群體（如：G5 爲五年級、G6 爲六年級、B 爲男生、G 爲女生）、最後一碼爲認知標的。舉例說明：「C7P(B)情」爲題本 C 第 7 題、爲建構反應題、DIF 拆題後的男生題，測「推論角色情緒」。

圖 2 顯示除 B1L1（G5）外，不分文本，所有試題的認知層次，從低至高皆爲「尋找和回憶」、「整合和解釋」及「批判和評鑑」。從表 6 可知三個認知層次的平均難度依序爲-0.89、0.32、1.20，達統計上顯著差異（p = .00），Scheffé 事後比較顯示高層次的平均難度皆高過低層次的平均難度，符合明細表中的預期階層關係。從圖 3 可知當學童能力達層次二之試題平均難度（0.32），其能力高過所有層次一試題的難度，也就是說其答對層次一題目的機率高過 .50；當學童能力達層次三之試題平均難度（1.20），其能力高過除 D7P 外所有層次二的試題，意即答對層次二絕大部份試題的機會高過一半。若從答對試題之勝算率言，能力達層次二試題平均難度的學童其勝算率是能力

達層次一試題平均難度學童的 3.35 倍（$e^{(0.32-(-0.89))} = 3.35$），而能力達層次三試題平均難度的學童其勝算率是能力達層次一及層次二之試題平均難度學童的 8.08 及 2.41 倍。

（三）與外在變項關係

1. 聚斂與區辨相關

表 7 爲效標樣本學童在本測驗之能力估計值與外在變項之相關，結果顯示與「中文閱讀理解測驗」之閱讀理解分測驗相關最高（$r = .68$）、與語言處理分測驗相關次之（$r = .63$），再其次爲與學童在校國語文學期成績之相關（$r = .54$），最後爲與數學學期成績之相關（$r = .52$）。此結果大致支持自編測驗具聚斂與區辨相關，但注意自編測驗與國語及與數學學期成績之相關的差異極小（國語及數學成績相關高達 .80，表示兩者可能僅反映學童之一般學習）。

表 7 「國小高年級閱讀理解測驗」與外在變項之相關 ($n = 165$)

中文閱讀理解測驗 [a]		學期成績(Z)	
語言處理分測驗	閱讀理解分測驗	國語文	數學
$.63^{***}$	$.68^{***}$	$.54^{***}$	$.52^{***}$

註：[a]「中文閱讀理解測驗」測七項次能力，其中「語言處理」含音韻處理、語法及語意，「閱讀理解」含理解文章基本事實、摘要重點大意、推論、比較分析等。

*** $p < 0.001$。

2. 年級、性別、國語文能力次群體之差異考驗

表 8 呈現不同次群體之閱讀理解平均數及差異考驗。結果顯示六年級學童的閱讀能力（0.0）顯著高於五年級學童（-.47）（$p < .001$），此結果符合兒童閱讀理解能力隨年齡增長而發展之理論，且與其他研究結果一致（如：林寶貴、錡寶香，2000），惟年級僅解釋閱讀理解約 6%的變異量。在性別方面，女生的閱讀能力（-0.15）顯著高於男生（-0.32）（$p < .001$），此亦與其他研究結果一致（如：柯華葳等人，2008），惟性別僅能解釋閱讀理解不到 1%的變異量。在國語文能力方面，較高能力組學童的閱讀表現皆高過較低能力組學童的表現（$p < .001$，Scheffé 事後比較：高 > 中 > 低），國語文成績可解釋閱讀理解 18%的變異量，此結果亦符合預期。

表 8 不同次群體閱讀理解之平均數及差異考驗

變項	水準	N	mean	sd	η^2	平均數差異考驗
年級	五	526	-0.47	0.83	.06	t（df = 1035.9）= 8.48^{***}
	六	526	0.00	0.94		
性別	男	534	-0.32	0.91	.01	t（df = 1050）= 2.96^{***}
	女	518	-0.15	0.92		
國語文能力 [a]	低 (L)	20	-1.13	0.61	.18	F（2，162）= 17.91^{***}
	中 (M)	129	-0.23	0.84		Scheffé事後比較：
	高 (H)	16	0.51	0.96		H > M > L

註：[a]依據效標樣本之國語文學期成績分組，低於-1 *sd* 爲「低能力」、介於± 1 *sd* 爲「中能力」、高於+1 *sd* 爲「高能力」。

*** $p < 0.001$。

二、信度分析

（一）受試者分隔信度

Rasch 模式之受試者分隔信度爲 .84，表示本測驗之試題能穩定區隔受試者在閱讀理解構念上的位置，換句話說，當我們給予受試者另一組測相同構念之試題時，受試者的排序會相對穩定。

（二）評分者信度

Rasch 多面相模式的結果顯示兩位評分者的嚴苛度分別爲 -.06、.06，勝算率 1.13（$e^{0.12}$）接近 1.0，表示評分者給分相當一致。

三、測驗分數的轉換與應用

以下應用 Rasch 客觀測量模式的特性說明分數的解釋與應用。爲方便測驗使用者，研究者將 logit 單位轉換爲百分制的量尺分數（0~100），轉換公式爲 $R = 50 + 10.87(l)$，其中 R 爲轉換後量尺分數、50 爲量尺分數的平均數、10.87 爲量尺分數的標準差（原始分數 0~40 分之全距約含 9.2 個 logits [0 分及滿分估計值約爲正、負 4.6 logits]，故在百分制下，1 個 logit 約爲 10.87 個單位的量尺分數[100/9.2=10.87]）、l 爲轉換前之學童能力或試題難度之 logit 值。

本文提供三種對照表：表 9「原始分數與 logit 值、量尺分數及 PR 值對照表」（受 DIF 試題拆題之影響，不同年級及性別次群體有其獨立的轉換表）、表 10「試題難度 logit 值及量尺分數對照表」及表 11「能力與試題難度之差異值及答題機率對照表」；讀者可同時參照圖 3 及表 3 來推論及診斷學童的閱讀理解表現。

（一）原始分數的轉換與比較

採用表 9 時，測驗使用者可以根據學童的原始分數，找出對應的量尺分數、能力估計 logit 值及次群體內的百分等級 PR 值。舉例說明，某位五年級男生的原始總分爲 30 分，查表得知其量尺分數爲 64、logit 值爲 1.29、PR 爲 99；若某六年級女生也得到相同的原始分數，則其量尺分數爲 63、logit 值爲 1.20、PR 爲 89。上述兩生原始分數相同，但量尺分數或 logit 值因受 DIF 拆題影響，有些微差異，惟此對答題機率的影響甚小；PR 值則因不同年級學童閱讀理解能力之差異，而有明顯不同。

我們可應用答題勝算率來比較學童的表現。試舉一例，若甲、乙兩生皆爲六年級女生，原始分數分別爲 20 及 30，我們可採下面兩種方法求勝算率：（1）以 logit 計算：甲、乙兩生 logit 值分別爲 0.0 及 1.20，在相同試題上，甲生答對試題的勝算率是乙生的 0.30 倍（$e^{(b1-b2)} = e^{(0.0-1.20)} = .30$），或乙生的勝算率是甲生的 3.32 倍（$e^{(b2-b1)} = e^{(1.20-0.0)} = 3.32$）；（2）以量尺分數計算：甲、乙兩生量尺分數分別爲 50 及 63，甲生低過乙生 13 分，相當於-1.196 個 logits（1 個 logit = 10.87 分，13/10.87 = 1.196），轉換爲勝算率約.30 倍（$e^{-1.196} = .30$），或乙生高過甲生 13 分，相當於 1.196 個 logits，轉換爲勝算率約 3.31 倍（$e^{1.196} = 3.31$）。

（二）推論及診斷受試者之閱讀理解表現

我們可推論不同年級及性別學童之整體表現，亦可透過推估個別學童在答對特定試題及認知層次上的機率，來診斷其閱讀理解的表現。

首先，從表 9 推論不同次群體學童的表現，要達到「尋找和回憶」、「整合和解釋」及「批判和評鑑」三層次的平均值，五年級男生的 PR 值需達 40、87 及 99，五年級女生 PR 值需達 30、81 及 96，六年級男生 PR 值需達 18、68 及 93，六年級女生 PR 值需達 14、58 及 89。

其次，對推論學童答對特定試題的機率做說明。若想知道原始分數爲 20 分之甲生（六年級女生）答對試題 D3 的機率。首先查表 9 得知甲生量尺分數爲 50（0.0 logit），查表 10 得知試題（D3L1）量尺分數爲 46（-0.34 logits），因甲生能力高過試題難度 4 分（能力減試題難度 = 50 - 46 = 4），

故查表 11 得知甲生答對該題的機率爲 .60。讀者亦可直接根據甲生能力與試題難度的 logit 差異值，求答對試題的機率： $p = e^{(.0-(-0.34))} / \left(1 + e^{(.0-(-0.34))}\right) = .58$（與查表得到的 .60 一致）。

表 9　原始分數與 logit 值、量尺分數及 PR 值對照表

原始分數	五年級男生（*n* = 265）			五年級女生（*n* = 261）			六年級男生（*n* = 267）			六年級女生（*n* = 257）		
	logit	量尺分數	PR	Logit	量尺分數	PR	logit	量尺分數	PR	logit	量尺分數	PR
0	-4.40	2	-	-4.59	0	-	-4.43	2	-	-4.61	0	-
1	-3.58	11	-	-3.72	10	-	-3.61	11	-	-3.75	9	-
2	-3.01	17	-	-3.12	16	-	-3.04	17	-	-3.15	16	-
3	-2.62	21	-	-2.71	20	-	-2.66	21	-	-2.74	20	-
4	-2.32	25	1	-2.39	24	1-	-2.35	24	1	-2.42	24	1-
5	-2.06	28	3	-2.12	27	1	-2.10	27	2	-2.16	26	1
6	-1.84	30	5	-1.89	29	2	-1.88	29	4	-1.93	29	2
7	-1.65	32	7	-1.69	32	5	-1.69	32	6	-1.73	31	2
8	-1.47	34	11	-1.51	34	8	-1.51	34	8	-1.55	33	4
9	-1.31	36	18	-1.34	35	12	-1.35	35	10	-1.38	35	7
10	-1.16	37	25	-1.19	37	17	-1.20	37	11	-1.23	37	9
11	-1.01	39	32	-1.04	39	24	-1.06	38	14	-1.09	38	11
12L1 平均	-0.88	40	40	-0.90	40	30	-0.92	40	18	-0.95	40	14
13	-0.75	42	46	-0.77	42	35	-0.79	41	22	-0.82	41	19
14	-0.62	43	50	-0.65	43	40	-0.67	43	26	-0.69	42	23
15	-0.50	45	55	-0.52	44	48	-0.55	44	30	-0.57	44	27
16	-0.38	46	60	-0.40	46	54	-0.43	45	35	-0.45	45	30
17	-0.26	47	64	-0.29	47	57	-0.31	47	40	-0.34	46	33
18	-0.15	48	68	-0.17	48	61	-0.20	48	44	-0.22	48	38
19	-0.03	50	72	-0.06	49	66	-0.09	49	48	-0.11	49	41
20	0.08	51	77	0.05	50	71	0.03	50	52	0.00	50	45
21	0.19	52	81	0.17	52	75	0.14	51	57	0.11	51	48
22	0.30	53	84	0.28	53	78	0.25	53	63	0.23	52	53
23L2 平均	0.42	55	87	0.39	54	81	0.36	54	68	0.34	54	58
24	0.53	56	89	0.51	55	84	0.48	55	74	0.45	55	64
25	0.65	57	91	0.62	57	87	0.59	56	79	0.57	56	68
26	0.77	58	93	0.74	58	89	0.71	58	81	0.69	57	73
27	0.89	60	95	0.87	59	92	0.84	59	84	0.81	59	78
28	1.02	61	97	0.99	61	93	0.96	60	87	0.93	60	81
29	1.15	62	98	1.13	62	94	1.09	62	90	1.07	62	85
30L3 平均	1.29	64	99	1.26	64	96	1.23	63	93	1.20	63	89
31	1.44	66	99	1.41	65	98	1.38	65	95	1.35	65	92
32	1.60	67	-	1.57	67	99	1.54	67	96	1.51	66	95
33	1.77	69	-	1.74	69	-	1.71	69	98	1.68	68	96
34	1.96	71	-	1.93	71	-	1.89	70	99	1.87	70	97
35	2.17	74	99+	2.14	73	-	2.11	73	99	2.08	73	99
36	2.42	76	-	2.39	76	-	2.35	75	99+	2.32	75	99
37	2.71	79	-	2.68	79	-	2.65	79	-	2.62	78	99
38	3.09	84	-	3.07	83	-	3.03	83	-	3.00	83	99+
39	3.65	90	-	3.62	89	-	3.58	89	-	3.55	89	-
40	4.46	98	-	4.43	98	-	4.39	98	-	4.36	97	-

註：原始分數標示 L1、L2、L3 的地方約爲「尋找和記憶」、「整合和解釋」及「批判和評鑑」三個認知層次的平均值。

表 10　試題 logit 值及量尺分數對照表

題號	logit	量尺分數	題號	Logit	量尺分數	題號	logit	量尺分數
A1L2	0.16	52	B7L2P.1	-0.11	49	C8L3P	1.06	61
A2L1	-0.44	45	B7L2P.2	0.91	60	D1L2	0.34	54
A3L2	-0.23	47	B8L2P(G5)	0.79	59	D2L1	-1.39	35
A4L1	-1.46	34	B8L2P(G6)	0.06	51	D3L1	-0.34	46
A5L2	0.42	54	B9L3P(G5)	1.54	67	D4L1	0.05	51
A6L2P	0.36	54	B9L3P(G6)	0.95	60	D5L1	0.05	50
A7L2P	0.55	56	C1L2	0.18	52	D6L1	0.36	54
A8L3P	0.82	59	C2L2	0.25	53	D7L2P	1.59	67
B1L1(G5)	0.17	52	C3L2	0.14	51	D8L3P	1.87	70
B1L1(G6)	-0.54	44	C4L1	-0.81	41	E1L1	-0.91	40
B2L1	-1.56	33	C5L1	-0.96	40	E2L1	-0.45	45
B3L2	0.25	53	C6L1(B)	-2.13	27	E3L2	-0.32	46
B4L2	0.11	51	C6L1(G)	-2.94	18	E4L1	-0.65	43
B5L2	-0.25	47	C7L2(B)	1.08	62	E5L2	-0.10	49
B6L2	0.63	57	C7L2(G)	0.34	54	E6L3P	0.97	60

註：題目 1-2 碼爲題本及題本內編號、3-4 碼爲認知層次，題號後註明 P 者爲建構反應題，題號句點後數字標示第 *k* 個閾值，(　)內標示 DIF 試題拆題後所隸屬的次群體（如：G5 爲五年級試題、G6 爲六年級試題、B 爲男生題、G 爲女生題）。

表 11　能力與試題難度之差異值及答題機率對照表

能力減試題難度	答題機率	能力減試題難度	答題機率	能力減試題難度	答題機率	能力減試題難度	答題機率
50	.99	24	.90	-2	.45	-28	.07
48	.99	22	.88	-4	.40	-30	.06
46	.99	19	.86	-6	.35	-33	.05
43	.98	17	.83	-9	.31	-35	.04
41	.98	15	.80	-11	.27	-37	.03
39	.97	13	.77	-13	.23	-39	.03
37	.97	11	.73	-15	.20	-41	.02
35	.96	9	.69	-17	.17	-43	.02
33	.95	6	.65	-19	.14	-46	.01
30	.94	4	.60	-22	.12	-48	.01
28	.93	2	.55	-24	.10	-50	.01
26	.92	0	.50	-26	.08		

最後，說明如何診斷學童的閱讀認知層次。以上述甲生爲例，其能力估計值爲 0 logit（量尺分數 50），從圖 3 可知，甲生能力與層次一最難三題[D5、D4、B1（G5）]的難度相當（約一半答對機率），其答對 D3（量尺分數 46）的機率 .60、答對最簡單一題 C6（G）（量尺分數 18）的機率約 .95，此結果顯示甲生已大致掌握處理文本明述訊息或表層文意理解的能力。但甲生仍未發展出整合與解釋文本訊息（層次二）的能力，甲生在答對層次二最簡單試題 E3（量尺分數 46）的機率雖達 .60，但在答對層次二平均難度左右試題[如：A6、C7（G）量尺分數 54]的機率僅約 .40，而答對層次二最難試題 D7（量尺分數 67）的機率更低至 .17，顯示甲生雖已開始發展層次二的認知能力，但仍無法跳脫文本提供的片斷訊息、觀點與細節，在一個較高及抽象的層次，進行統整與解釋，此時，甲生也勢必難將閱讀新知應用於生活中。最後，甲生在回答層次三試題的機率最高僅 .30 左右（層次三最簡單一題 A8 的量尺分數爲 59），顯示甲生尚未開始發展層次三的認知能力，無法從多元觀點批判與評鑑作者動機、選擇語言的目的、使用手法/技巧等。

討論

本研究根據 NAEP 閱讀評量架構，以 Rasch 客觀測量模式建置適合國小高年級一般學童的文本閱讀理解測驗，並建立信、效度憑證。結果顯示：（1）在測驗內容方面，專家判斷結果支持測驗內容與閱讀理解構念具關聯性且爲代表性樣本；（2）在構念向度上，個別試題契合度指標、殘差主成份分析、性別及年級 DIF 分析結果支持本測驗所測閱讀理解構念爲單向度，學童及試題可置於同一量尺上相互比較；（3）在建構反應題之等級結構上，經閾值分析選擇適切的等級數；（4）在試題難度適切性上，試題難度分配廣泛且與學童能力相對應，可測不同能力之學童；（5）在文本及題型難度上，不同文本試題之難度分配雖略不同，但無顯著差異；建構反應題之難度顯著高於選擇題之難度，分別可測高/中、低/中能力的學童；（6）在閱讀理解認知層次上，難度從易至難依序爲尋找和回憶、整合和解釋、批判和評鑑，符合預期的階層關係；（7）在本測驗與其他測量之相關上，從高至低依序爲：「中文閱讀理解測驗」、國語文學期成績、及數學學期成績，大致支持本測驗具聚斂與區辨相關；（8）在學童閱讀理解表現之差異上，六年級表現優於五年級、女生表現優於男生、在校語文能力高的學童表現優於語文能力低之學童，支持本測驗能區辨不同年級、性別、語文能力學童的閱讀理解表現；（9）在信度方面，受試者分隔信度係數佳且評分者給分一致性高。

一、「國小高年級閱讀理解測驗」之應用

本測驗適用國小五、六年級一般學童，可用於推論、比較並診斷學童之文本閱讀理解能力。測驗結果雖然可以從常模參照及效標參照雙重觀點解釋學童的表現，但從效標參照的角度診斷學童在個別試題及閱讀理解層次上的表現，能獲得更深入的訊息，對教學助益大。

從常模參照角度，我們可以透過 PR 值來瞭解並比較學童在不同次群體內的相對表現，亦可透過答題勝算率來比較同一次群體或不同次群體學童的表現。從效標參照角度，可以藉由 Wright Map（圖 2、3）、試題認知標的（表 3）、各類分數轉換表（表 9 至表 11）來推估學童在個別試題及不同認知層次上的答題機率，進而推論其可能處在的認知發展層次。之後，教師可針對學童所處在的認知發展層次，設計接近其閱讀理解能力的教學活動。譬如：若我們知道某生的發展仍不及層次一「尋找和回憶」的平均值，此時教師應加強該生有關層次一的各項標的行爲（如：找出段落內/跨段落之特定人/事/時/地/物的訊息、找出支持細節、做段落內/跨段落的簡單推論、確認比喻性言語等），而非強調高層次的認知標的；但若學生已處於發展層次三「批判與評鑑」，此時教師就應加強層次三的相關認知標的（如：評斷文章整體品質、評斷作者的寫作動機、作者使用語言/技巧的有效性與適切性、評斷立論的可信度與充份性等），而非重複提供層次一的練習。

由於本研究結果也顯示學童的表現或發展受語文能力、年級及性別的影響，故現場教師在規畫閱讀教學時，也宜考慮這些因素。

二、Rasch 模式之應用

本研究顯示當我們以 Rasch 模式建構一個具客觀測量特性量尺的同時，也建立了這個測量的各項信、效度憑證。在編製測驗上，我們可以檢視構念的向度性、選擇廣度與精確性兼顧的試題、同時納入適合測不同認知層次之各類型題目、選擇適合各個開放題的等級數、檢視評分者在開放題上給分的一致性、檢驗試題是否因不同群體而測到不同構念（DIF）並做處理等。一但具客觀測量特性的量尺被建立，此時，量尺具真正等距的特性，試題及受試者可置於同一構念向度上相互比較，我們不僅可以推論、診斷及比較個別學童的表現，更可嘗試建立學童在閱讀理解構念上的

發展層次，提供更深入的診斷訊息。因此，Rasch 模式不僅爲閱讀理解測驗的編製，更爲閱讀理解的相關研究（如：認知發展階段）提供了更廣泛的發展空間。

三、後續研究之建議

根據研究結果與限制，研究者對後續研究提出一些建議。首先，在測驗編製方面，因編題上的困難，本測驗並未以選擇題評量「批判和評鑑」層次，建議後續研究可開發以選擇題測量較長文本之高層次的文本閱讀理解能力，並檢視結果是否與本研究結果一致。其次，針對閱讀理解的計量特性，本研究結果雖強烈暗示學童的閱讀理解構念呈現「尋找和回憶」、「整合和解釋」及「批判和評鑑」三個發展層次，以及不同年級的學童在這三個發展層次上有所不同，惟因本研究的主題並不在探討閱讀理解的發展階層，故無法得知此階層性是否具跨不同教學、年級、文本性質及其他因素的恆定性，建議未來研究可針對此，進行後續的探討。另外，本研究也發現，特定性質（或內容）的文本試題可能對不同次群體（如：年級、性別）測到不同構念，意即部份試題呈現 DIF 現象，建議未來研究可針對造成「文本閱讀理解 DIF」的因素及解決方法（如：本研究並未刪除試題，而是以拆題[即試題連結]的方式解決 DIF 問題），做更多的探討。最後，從教學研究的角度言，高層次閱讀認知能力通常需藉由建構反應題測得，惟我國教學現場，除學生不習慣書寫此類型題目（相當比例的學童作答意願低，填寫「不知道」或「空白」）外，教師礙於時間不足或與學校評量方式不同，甚少採用此類型的評量，導致教師可能在編製此類型試題、計分規準及評分上均會遭遇困難；建議未來的教學研究可針對現場教師，舉辦相關的研習或工作坊。

參考文獻

王文中（1996）：幾個有關 Rasch 測量模式的爭議。**教育與心理研究，19**，1-26。[Wang, W. C. (1996). Some controversial issues about the Rasch measurement model. *Journal of Education & Psychology, 19*, 1-26.]

邱上真、洪碧霞（1996）：**國語文低成就學生閱讀表現之追蹤研究（I）**。國科會專題研究報告（編號：NSC84-2421-H-017-00-F5）。[Chiu, S. C., & Hung, P. H. (1997). *A longitudinal study of Chinese language low achievers on reading performance* (I). NSC report (NSC84-2421-H-017-00-F5).]

邱上真、洪碧霞（1997）：**國語文低成就學生閱讀表現之追蹤研究（II）**。國科會專題研究報告（編號：NSC86-2413-H-017-002-F5）。[Chiu, S. C., & Hung, P. H. (1997). *A longitudinal study of Chinese language low achievers on reading performance* (II). NSC report (NSC86-2413-H-017-002-F5).]

林建平（1994）：**整合學習策略與動機的訓練方案對國小閱讀理解困難兒童的輔導效果。**國立台灣師範大學教育心理與輔導研究所博士論文。[Lin, C. P. (1994). *Effects of a training program combining learning strategies and motivation on elementary school children with reading difficulties* (Unpublished doctoral dissertation). National Taiwan Normal University, Taipei, Taiwan.]

林寶貴、楊慧敏、許秀英（1995）：中華國語文能力測驗之編製及相關因素之研究。**特殊教育研究學刊，12**，1-24。[Lin, B. G., Yang, W. M., & Hsu, S. Y. (1995). A study of the Chinese language ability test and its correlated factors in Taiwan, R. O. C. *Bulletin of Special Education, 12*, 1-24.]

林寶貴、錡寶香（2000）：中文閱讀理解測驗之編製。**特殊教育研究學刊，19**，79-104。[Lin, B. G., & Chi, P. H. (2000). The development of test of reading comprehension. *Bulletin of Special Education, 19*, 79-104.]

胡永崇（1995）：**後設認知策略教學對國小閱讀障礙學童閱讀理解成敗之研究。**國立彰化師範大學特殊教育研究所博士論文。[Hu, Y. C. (1995). *Examining the effects of metacognitive strategy instruction on reading comprehension of elementary school children with reading disabilities* (Unpublished doctoral dissertation). National Changhua University of Education, Changhua, Taiwan.]

胡志偉、顏乃欣（1995）：中文字的心理歷程。載於曾進興（主編），**語言病理學基礎**（30-76）。台北：心理。[Hue, C. W., & Yen, N. S. (1995). The psychological process of comprehending Chinese characters. In Tseng C. H. (Ed.). *Foundation of language pathology* (pp. 30-76). Taipei, Taiwan: Psychological Publishing Co.]

柯華葳(1999)：閱讀理解困難篩選測驗。**中國測驗學會測驗年刊，46**，1-11。[Ko, H. W. (1999). Reading comprehension difficulties screening test. *Psychological Testing, 46*, 1-11.]

柯華葳、詹益綾、張建妤、游婷雅（2008）：**台灣四年級學生閱讀素養-PIRLS 2006 報告**。取自PIRLS2006 國際報告：http://lrn.ncu.edu.tw/pirls/ pirls%202006%20report.html，2009 年 2 月 25 日。[Ko, H. W., Chan, Y. L., Chang, C. Y., & Yu, T. Y. (2008). *PIRLS 2006 National Report: Reading literacy study of fourth grade Taiwanese students.* Retrieved February 25, 2009, from http://lrn.ncu.edu.tw/pirls/ pirls%202006%20report.htm]

高雄師範大學工業科技教育學系（2009）：**中文文章適讀性線上分析系統**。取自中文文章適讀性線上分析系統：http://140.127.45.25/readingability/Analyze，2009 年 4 月 5 日。[Department of Industrial Technology, National Kaohsiung Normal University (2009). *Analyzing system for readability of Chinese articles.* Retrieved April 5, 2009, from http://140.127.45.25/readingability/Analyze]

教育部（2008）：**國民教育社群網-97 年課程綱要**。取自國民教育司：http://teach.eje.edu.tw/ 9CC2/9cc_97.php，2009 年 4 月 5 日。[Minstery of Education (2008). *Taiwan elementary and secondary educator community - curriculum guidelines.* Retrieved April 5, 2009, from http://teach.eje.edu.tw/ 9CC2/9cc_97.php]

教育部（2009，10 月）：**全方位的國民中小學閱讀推動政策**。取自教育部全球資訊網：http://www.edu.tw/PDA/news.aspx?news_sn=2798&pages=16&unit_sn =15，2009 年 10 月 8 日。[Minstery of Education (2009, October). *The full range promotion of elementary and junior high school's reading policies.* Retrieved October 8, 2009, from http://www.edu.tw/PDA news.aspx?news_sn=2798&pages=16&unit_sn =15]

曾志朗（1991）：華語文的心理學研究：本土化的沉思。載於楊中芳、高尚仁（主編）：**中國人、中國心--發展與教學篇**（539-582）。臺北：遠流。[Tzeng, J. L. (1991). Psychological studies of Chinese language and literacy: Contemplation on localization. In Yang C. F. & Kao H. S. R. (Eds.), *Zhongguo ren, zhongguo xin: Development and Instruction* (pp. 539-582). Taipei, Taiwan: Yuan-Liou.]

張麗麗、羅素貞（2011）：Rasch 多向度模式檢核「國小數學問題解決態度量表」（MPSAS）之心理計量特性。**教育與心理研究，34**（3），153-185。[Chang, L., & Lo, S. J. (2011). Using the multidimensional Rasch model to examine the psychometric properties of Mathematics Problem-Solving Attitude Scale (MPSAS). *Journal of Education & Psychology, 34*(3), 153-185.]

董宜俐（2003）：**國小六年級學童中文閱讀理解測驗編製研究。**國立台中師範學院教育測驗統計所碩士論文。[Tung, Y. L. (2003). *The development of the Chinese reading comprehension test for the 6th graders* (Unpublished master's thesis). National Taichung University, Taichung, Taiwan.]

Andrich, D. (1988). *Rasch models for measurement.* Beverley Hills, CA: Sage.

Andrich, D., & Hagquist, C. (2004, Jan). *Detection of differential item functioning using analysis of variance.* Paper presented at the Second International Conference on Measurement in Health, Education, Psychology and Marketing: Developments with Rasch Models. Perth, Australia: Murdoch University.

Andrich, D., Sheridan, B., & Luo, G. (2007). *RUMM2020*. Perth, Australia: RUMM Laboratory.

Engelhard, G. Jr. (1993). What is the attenuation paradox? *Rasch Measurement Transactions, 6*(4), 257.

Holland, P. W., & Thayer, D. T. (1988). Differential item performance and the Mantel-Haenszel procedure. In H. Wainer & H. I. Braun (Eds.), *Test validity* (pp. 129-145). Hillsdale, NJ: Erlbaum.

Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychology Review, 95*, 163-182.

Linacre, J. M. (2004). Optimizing rating scale category effectiveness. In E. V. Smith & R. M. Smith (Eds.), *Introduction to Rasch measurement: Theory, models and applications.* (pp. 258-278). Maple Grove, Minnesota: JAM Press.

Linacre, J. M. (2005). *A user's guide to WINSTEPS: Rasch-model computer programs.* Chicago, IL: Winsteps.com.

Master, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*(2), 149-174.

NAGB (National Assessment Governing Board) (2008). *Reading assessment and item specifications for the 2009 National Assessment of Educational Progress*. Washington, DC: American Institutes for Research.

OECD (2007). *PISA 2006: Science competencies for tomorrow's world (Executive summary)*. Retrieved from OECD: http://www.oecd.org, 10, 8, 2009.

OECD (2010). *PISA 2009 results: Learning to learn.* Retrieved from OECD: http://www.oecd. org, 7, 8, 2011.

Rasch, G. (1980). *Probabilistic models for some intelligence and attainment tests* (revised and expanded ed.). Chicago, IL: University of Chicago Press. (Original work published 1960)

Singh, J. (2004). Tackling measurement problems with item response theory: Principles, characteristics, and assessment, with an illustrative example. *Journal of Business Research, 57*, 284-208.

Smith, R. M., & Miao, C. Y. (1994). Assessing unidimensionality for Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2*,* pp. 316-327). Norwood, NJ: Ablex Publishing Corporation.

Smith, E. V. (2001). Evidence for the reliability of measurement and the validity of measure interpretation: A Rasch measurement perspective. *Journal of Applied Measurement, 3*, 205-231.

Smith, E. V. (2005). Effect of item redundancy on Rasch item and person estimates. *Journal of Applied Measurement, 6*(2), 147-163.

Swaby, B. E. R. (1989). *Diagnosis and correction of reading difficulties*. Boston, MA: Allyn and Bacon.

van der Linden, W. (1994). Fundamental measurement and the fundamentals of Rasch measurement. In M. Wilson (Ed.), *Objective measurement: Theory into practice* (Vol. 2, pp. 3-24). Norwood, NJ: Ablex.

Wright, B. D. (1996). Local dependency, correlations and principal components. *Rasch Measurement Transactions, 10*(3), 509-511.

Wright, B. D., & Linacre, J. M. (1994). Reasonable mean-square fit values. *Rasch Measurement Transactions, 8*(3), 370.

Yen, W. M. (1993). Scaling performance assessments: Strategies for managing local item dependence. *Journal of Educational Measurement, 30*(3), 187-213.

收 稿 日 期：2012 年 04 月 01 日
一稿修訂日期：2012 年 09 月 03 日
二稿修訂日期：2012 年 10 月 11 日
三稿修訂日期：2012 年 11 月 28 日
接受刊登日期：2012 年 11 月 28 日

Bulletin of Educational Psychology, 2013, 45(1), 39-61
National Taiwan Normal University, Taipei, Taiwan, R.O.C.

The Development of Reading Comprehension Test for 5th and 6th Graders Using the Rasch Model

I-Chun Lin
Kaohsiung City
Dingjin Elementary School

Lily Chang
I-Chung Lu
Department of Educational Psychology and Counseling
National Pingtung University of Education

The Reading Comprehension Test, based on NAEP reading framework, was developed for 5th and 6th grade students using the Rasch measurement model. The participants of this study were 1,052 5th and 6th graders sampled from Kaohsiung and Pingtung counties using stratified random sampling procedures. The 39-item test, including both multiple-choice and construct-response items, measures three cognitive levels (i.e., locate/recall, integrate/interpret, and critique/evaluate) of literary and informational texts. Results show: a) the underlying trait defined by the reading comprehension test holds the characteristics of an objective measurement (i.e., the construct measured by the test is unidimensional and invariant across genders and grades); b) the rating scales for construct-response items are appropriate in terms of their order and distance; c) items spread reasonably well along the latent continuum and are aligned with various ability levels; d) the three reading cognitive levels show hierarchical structure as expected; e) the test converges and discriminates various measures appropriately; f) the test is able to discriminate group differences by gender, grade, and ability level; and g) the test has satisfactory person separation and rater reliability. Discussions and suggestions regarding test application, the application of the Rasch model, and future research are provided.

KEY WORDS: NAEP, rasch model, reading comprehension test