國立臺灣師範大學教育心理與輔導學系
教育心理學報，2021，53 卷，2 期，285–306 頁
https://doi.org/10.6251/BEP.202112_53(2).0002

中文閱讀能力適性診斷評量編製研究*

李奕璇　　周業太　　宋曜廷

國立臺灣師範大學
心理與教育測驗研究發展中心

國立臺灣師範大學
教育心理與輔導學系

本研究建置了一套可同時評測整體閱讀理解能力，並診斷出字詞辨識、表層文意理解、文意統整、推論理解、分析評鑑之閱讀細項技能程度的中文閱讀能力適性診斷評量系統。此系統的適用對象為二至十二年級學生，為相關領域第一套可橫跨多個學習階段的中文閱讀測驗。本測驗透過現代測驗理論技術估算試題難度與學生能力參數，另藉由題庫建置、常模建置等作法，利用電腦化適性測驗的型式施測，可快速且精確地定位學生的閱讀能力程度，並長期追蹤能力變化情形。分析結果指出本測驗具備良好的重測信度、效標關聯效度、條件化信度、與 IRT 效度，顯示本測驗具有優秀的品質，可有效且穩定地評量學生的中文閱讀能力。

關鍵詞：閱讀理解、電腦化適性測驗、診斷、中文閱讀能力

* 1. 通訊作者：宋曜廷，sungtc@ntnu.edu.tw。
2. 本研究獲教育部高等教育深耕計畫之國立臺灣師範大學華語文與科技研究中心經費補助，特此致謝。

閱讀理解能力（reading comprehension）是所有學科學習的重要基礎，也是人與人溝通和互動的基本能力，更是終身學習的關鍵素養。因此，近年來在閱讀相關領域的學術研究蓬勃發展，如，閱讀教學重要成分（National Reading Panel, 2000）的整合、閱讀策略教學的課程設計（陳茹玲、宋曜廷等人，2017；謝進昌，2019）、閱讀歷程與閱讀表徵（沈欣怡、蘇宜芬，2011）的探討、數位教學或學習工具的開發（Crossley & McNamara, 2017; Hsiung et al., 2017; Lan et al., 2017; Liao et al., 2020; Rosita Cecilia et al., 2016; Sung et al., 2019; Yang et al., 2020）、特定型式評量的研發與驗證（林小慧、曾玉村，2017）、文本可讀性分析（陳茹玲、曾厚強等人，2017；Hong et al., 2020）、抑或是文本難度分級與推薦（陳昭珍等人，2020；Tseng et al., 2019）等皆有豐碩的研發成果，可見學術領域與教學現場對閱讀理解能力的重視。但在探討閱讀理解能力的應用之前，首先應掌握的便是如何準確地測量學生的能力程度。這也是國際大型評比測驗，如，促進國際閱讀素養研究（Progress in International Reading Literacy Study，簡稱 PIRS）、學生能力國際評量計畫（Program for International Student Assessment，簡稱 PISA) 等，皆將閱讀能力納入評量指標的原因之一。

有鑑於閱讀的重要性，在教育研究的領域中，因應不同的使用需求，已開發出各式閱讀能力測驗。然而，綜觀現有閱讀測驗，仍可發現諸多不足之處。在國內，常見的閱讀測驗有：國小學童中文閱讀理解測驗（王木榮、董宜俐，2006）、閱讀理解量表（張世彗，2014）、國小句型理解測驗（張祐瑄、蘇宜芬，2014）、中文閱讀理解測驗（林寶貴、錡寶香，2000）、閱讀理解成長測驗（蘇宜芬等人，2018）等。這些評量工具雖透過嚴謹的測驗編製技術研發而成，亦具有不錯的信度與效度，然目標對象卻都較聚焦於特定的年級（如，國小）或是族群（如，閱讀障礙學生），無法廣泛地適用於所有學生。

此外，常見的閱讀測驗大多使用紙本進行施測。紙本測驗雖可在課堂上直接施測，但往往須要耗費時間、人力與物力進行印製、分卷、批閱、計分整理等工作，從測驗完到取得分數，以至於後續的分數應用曠日廢時。且紙本評量的測驗題型和題材有許多限制，大多為單一、靜態的文字搭配選擇題，然而要真實評量學生的閱讀理解能力，這些材料遠遠不足。學生在生活當中，除了傳統的連續性文本以外，還會碰到許多非連續文本，例如宣導手冊、圖表說明，考驗學生的圖文整合與文意闡釋等閱讀理解能力，甚至以網頁形式編排的超連結動態文本，更是現時學生們閱讀的主體。如要完整地評量學生的閱讀能力，絕不能忽略這些題材，但紙本測驗卻難以納入這些材料，唯有採用電腦化測驗方能融合多元的素材，真實且全面地評量學生的閱讀技能。

其次，要準確地測量學生的閱讀能力，傳統上需要透過大量的閱讀測驗試題才能達成，但是過量的題目卻會造成學生作答上的疲乏，降低測驗的信度與效度；同時，測驗題目的命題也須耗費大量的人力及經費成本，但這些耗費大量資源研發的測驗題目，往往卻由於試題曝光的因素，僅能施測一次。因此，致使市面上難有相關工具能有效且長期地透過多次施測，追蹤學生的閱讀能力發展。

綜觀現有閱讀能力測驗，可發現市面上並沒有可長期持續應用於課室評量的中文閱讀能力評量。本研究的目的為：一、改善現有閱讀理解測驗工具的限制，利用先進的評量技術，比照國際測驗（如，TOEFL、GRE 等），採用電腦化適性測驗（computerized adaptive testing）的形式，研發出一套可全面診斷學生閱讀能力表現的評量工具，稱為「中文閱讀能力適性診斷評量」（Diagnostic Assessment of Chinese Competence，簡稱 DACC）。二、探討 DACC 的信度與效度。

（一）國外閱讀理解測驗現況

國際間有多項標準化測驗皆將閱讀理解能力列入其測驗架構中，規模較大者有：PIRLS、PISA，以及美國的全國教育進步測驗（National Assessment of Educational Progress，簡稱 NAEP）。

PIRLS 為國際教育成就評鑑協會（The International Association for the Evaluation of Educational Achievement，簡稱 IEA）所主辦的國際性評比，評量對象為國小四年級學生。此評比每五年舉辦一次，以最近一次測驗 2016 年為例，共有 48 個國家／地區參與。臺灣在 2016 年當屆共有 150 所學校，4,326 位學生參加此測驗。PIRLS 測驗的目的為瞭解學生文學賞析及獲得與使用資訊的能力，並透過四個閱讀理解歷程檢視學生的閱讀理解能力。此四個歷程分別是：直接提取（focus on and retrieve explicitly stated information）、直接推論（make straightforward inferences）、詮釋整合（interpret and integrate ideas and information）、比較評估（examine and evaluate content, language, and textual

elements）（柯華葳等人，2017）。PIRLS 2016 共有 12 篇文章，包括兩種文體——故事體與說明文：故事體用來瞭解學生的文學賞析能力，說明文評量學生獲得與使用資訊的能力，測驗題型包含單一選擇題與問答題（柯華葳等人，2017）。

另一項廣為人知的跨國大型閱讀能力評比為 PISA。此測驗由經濟合作暨發展組織（The Organization for Economic Co-operation and Development, OECD）發起，每隔三年對參與國家之十五歲學生依標準化程序，進行抽樣與檢測。PISA 的評量目的為瞭解即將受完義務教育的學生是否已準備好面對未來的挑戰，是否有能力分析、推理、及溝通他們的觀點。它不僅測量學生所學的知識，更評量學生能否應用所學至不熟悉的情境。PISA 在閱讀素養上所測量的閱讀歷程包括：1. 擷取與檢索（access and retrieve）：偵測或辨識出問題的一個或多個重要元素，找出文中直接陳述的資訊；2. 統整與解釋（integrate and interpret）：瞭解文本各部分的關係，例如因果關係、分類與舉例、比較等，並從未明事物中建構意義，如，文本訊息的演繹、推論；3. 省思與評鑑（reflect and evaluate）：連結文本資訊與外在知識，瞭解文本目的，評鑑文本的品質與適切性（Organization for Economic Co-operation and Development, 2010）。文章類型則包括記敘文、敘事文、說明文、論說文、指引或說明書等。

而 NAEP 為美國最大型的標準化測驗，自 1969 年起便已持續實施至今。與前述兩項大型測驗相同，評量的形式為透過既定的抽樣程序，對學生進行施測，對象是四年級、八年級與十二年級的學生。在閱讀測驗上，NAEP 所評採用的文本類型分為兩種：文學性文本（literary text）以及知識性文本（informational text）。評量的閱讀讀理解成分包括：1. 找出文本中的特定訊息（locate/recall）；2. 統整與解釋文本訊息（integrate/interpret）；3. 批判與評估文本訊息（critique/evaluate）（National Center for Education Statistics [NCES], 2009）。以最近一次 2019 年的閱讀評量施測結果而言，約有 150,600 位四年級學生（分布於 8,300 所學校）以及 143,100 為八年級學生（分布於 6,950 所學校）參與。

綜合檢視 PIRLS、PISA、NAEP 三項大型測驗所測量的閱讀理解成分，可發現三者均包含：擷取文本中清楚敘述的特定訊息，統整與解釋文意，及評估、反思、與批判等核心閱讀技能，故本測驗亦參考相關內涵建置測驗向度。

（二）國內閱讀理解測驗現況

而在國內的教學現場上，閱讀能力的培養亦受到許多教師與教學研究者的重視，但在評估學生的閱讀理解能力時，往往所採用的工具多為自行研發的測驗試題。這些測驗試題大多與課堂所使用的教學素材（例如，教科書的選文或補充文章等）相關，因此評量內涵常常需要呼應教學重點，而偏重於學科知識的檢核，像是字詞讀音的正確性、詞彙定義的記憶等等。且，閱讀理解能力雖為跨學科的核心素養，但在大部分的情況下，僅在國文課觸及閱讀技能的養成，故閱讀的材料也都以文學性的篇章為主，而缺乏不同類型的文本。

除了課堂閱讀測驗以外，其他國內常見的閱讀測驗有：國小學童中文閱讀理解測驗（王木榮、董宜俐，2006）、閱讀理解量表（張世彗，2014）、國小句型理解測驗（張祐瑄、蘇宜芬，2014）、中文閱讀理解測驗（林寶貴、錡寶香，2000）、閱讀理解成長測驗（蘇宜芬等人，2018）等。

「國小學童中文閱讀理解測驗」由王木榮與董宜俐於 2006 年編製，測驗內容針對國小六年級學童現階段閱讀理解能力加以評量，並提供閱讀理解教學的相關策略。檢測的能力為字義理解、文本理解、推論理解、摘要能力、布題能力。常模樣本為 335 位國小六年級學童。此測驗為紙本形式的固定題本測驗，測驗時間為 60 分鐘。而「閱讀理解量表」則為另外一項以國小二、三年級學生為適用對象的閱讀評量工具。其內容包含字詞釋意（10 題）、命題組合（7 題）、句子理解（8 題）及閱讀測驗（8 題），共計 33 題。此測驗預試人數 169 人，取樣對象皆為臺北市學生；正式樣本 1,039 人，來自臺北市八所國小、新北市 6 所國小。

相較於上述兩項適用對象為限定年級之測驗，亦有目標對象相對廣泛的評量工具，例如，「國小句型理解測驗」以及「中文閱讀理解測驗」。「國小句型理解測驗」由張祐瑄與蘇宜芬（2014）研發，以複句與關聯詞理論歸納出國小階段常用的九種複句（並列、連貫、遞進、轉折、假設、目的、因果、選擇、條件），並藉以進行上述閱讀技能的評估，篩選句型理解表現落後的學生及其落後的

句型類型進行補救教學。此測驗分為三個版本，二至六年級版本共 36 題，每種複句 4 題，用以初步篩選二至六年級句型理解表現落後的學生。二至三年級版本共 54 題，每種複句 6 題，以瞭解二至三年級學生的句型理解表現與在各種複句句型的掌握程度。四至六年級版本共 72 題，每種複句八題，測驗目的同樣為評估四至六年級學生的句型理解表現與在各種複句句型的掌握程度。測驗題型為四選一的單選題，測驗建有常模（臺灣地區 3,558 位國小二年級至六年級學生），可將測驗總分對照年級及地區百分等級。相較於前述兩個測驗，此測驗的常模樣本較多，測驗適用對象也較為廣泛。

另一方面，「中文閱讀理解測驗」（林寶貴、錡寶香，2000）為國內較早研發完成的標準化閱讀理解能力測驗，因此較為教育研究者所熟知。此測驗的適用對象亦為國小二至六年級學童，測驗目的為篩選在閱讀理解上有困難的學生，或是做為探討身心障礙學生閱讀理解能力之用。此測驗設計六篇故事類記敘文與六篇說明文，並在每篇文章下命製音韻處理、語法、語意、理解文章基本事實、摘要重點大意、推論、比較分析等題目。測驗的常模樣本取自臺灣北、中、南、東四區，共 30 所學校 773 位學童。「中文閱讀理解測驗」雖可適用較多年級的學生，但僅有單一版本，較難顧及不同年級程度的需求。

在所有測驗中，「閱讀理解成長測驗」（蘇宜芬等人，2018）為近期發布的研發成果，其適用對象為四至六年級學生。此測驗具有上述其他評量工具缺乏之優點，由於以多複本的形式建構，「閱讀理解成長測驗」可用來追蹤學生的閱讀理解能力發展情形。測驗題目型式為題組題。每套複本皆有 4 個題組，每個題組都有 10 道子題。4 個題組的文章分別是短記敘文、短說明文、長記敘文、長說明文。題目所測量的閱讀理解成份為字彙觸接、字面理解、摘取大意、推論理解。此測驗透過分階段預試等方式，施測 3,600 位學生，用以建構常模，但值得注意的是，此測驗在跨年級的題本間並無設計共同題進行等化，故不同年級版本的分數無法進行比較。

其他常見的閱讀測驗尚有柯華葳與不同研究者所研發的一系列閱讀診斷測驗，包含：閱讀理解困難篩選測驗（柯華葳，1999）、閱讀理解篩選測驗（柯華葳、詹益綾，2006）等。但由於此系列測驗的評量目標與內涵相似，對象則以特殊生為主，故在此不再分批贅述。下表為針對國內常見的閱讀能力相關測驗所進行的綜合整理：

表 1
國內常見閱讀測驗內容

測驗名稱	對象	形式	評量技能
國小學童中文閱讀理解測驗	六年級	單一版本	字義理解、文 本理解、推論理解、摘要能力、布題能力
閱讀理解量表	二、三年級	單一版本	字詞釋意、命題組合、句子理解、閱讀測驗
國小句型理解測驗	二至六年級	不同年級不同版本	九種複句（並列、連貫、遞進、轉折、假設、目的、因果、選擇、條件）
中文閱讀理解測驗	二至六年級	單一版本	音韻處理、語法、語意、理解文章基本事實、摘要重點大意、推論、比較分析
閱讀理解成長測驗	四至六年級	各年級皆有多複本	字彙觸接、字面理解、摘取大意、推論理解

由上述文獻可發現，市面上的閱讀評量工具有諸多限制，例如，這些測驗大多側重於特定年齡，或是以特定目的（如，評量學科能力、篩選閱讀障礙學生等）建構；且幾乎皆為紙本、無複本的測驗。這樣的測驗形式與特性致使教育工作者難以長期追蹤學生閱讀能力的變化情形，也沒有標準化的測驗結果能全面評估跨族群（例如，不同學校之間、不同區域之間）的閱讀理解表現。而大型標準化測驗如 PIRLS 或 PISA，雖有嚴謹的流程設計，卻都是以抽測的方式，僅對部分學生施測，亦不提供個別學生的測驗結果，因此也無法將相關結果回饋至日常教學使用。

（三）閱讀理論文獻探討

為了有效且完整地評量學生的閱讀能力，在研發以學齡學生為對象之閱讀測驗上，實有必要結合閱讀發展歷程理論，針對此階段學生的各項發展進程進行精準的評估與定位。在諸多學者所提出的閱讀發展理論中，Chall（1996）所提出的閱讀發展階段論將國小一年級至高中三年級有關的閱讀階段定義為：初始閱讀階段（initial reading or decoding stage），流暢閱讀階段（confirmation, fluency, ungluing from print），閱讀新知階段（reading for the new），及多元觀點階段（multiple viewpoints）。此四個階段的發展重點為：1.「初始閱讀階段」：約六至七 歲，相當於一至二年級。此時的發展重點在識字，學習者能覺察文字與讀音之間的對應關係。此階段在閱讀時容易發生認字上的錯誤；2.「流暢閱讀階段」：約 7 至 8 歲，相當於二至三年級。此時的發展重點在識字技能的自動化，亦即能流暢地閱讀。此階段比較能夠精確地辨識文字，以及建構文字與意義間的連結；3.「閱讀新知階段」：約 9 至 14 歲，相當於四至八年級。此時的發展重點為透過閱讀學習新知。此階段學生不僅藉由大量閱讀吸收知識，也開始發展閱讀策略；4.「多元觀點階段」：約 14 至 18 歲，相當於國中的八年級至高中三年級。此時的發展重點為豐富觀點。此階段的學生有能力理解內容複雜、觀點多元的文章，並能對內容分析，形成初步的批判。

而 van den Broek 等人（2005）根據過去認知心理學及心理語言學的研究，亦針對理解能力的發展提出看法。首先，他們認為閱讀理解的核心歷程是：讀者透過推論歷程，辨識、發現文本中不同部分之間的關係，以及文本中的這些部分與讀者先備知識的關係，然後在心中建構一個具連貫性的心理表徵。在「辨識關係」（identify relation）上，不同年齡的讀者有些發展上的差異。例如：年齡小的學生能發現具體事件之間的關係；年齡大的學生則也能發覺抽象事件之間的關係。年齡小的學生能發現外在事件（external events）之間的關係；年齡大的學生則也能發覺內在事件（internal events，如，主角的目標、情緒）之間的關係。年齡小的學生能發現個別事件之間的關係；年齡大的學生則也能發覺一組事件與另一組事件（如，情節）之間的關係。

若依推論的類別來看閱讀理解的發展，則 van den Broek 等人（2005）的觀點如下。首先，兒童剛開始能夠發現的關係是文本中鄰近事件之間，具體的、而且比較跟身體有關的關係。例如：文本中有兩個鄰近的句子，「小明把香蕉皮丟到地上」，「小美滑倒了」，則兒童通常會推論「小美踩到了小明丟的香蕉皮」。其次，兒童會開始能夠發現文本中距離較遠之事件的關係。第三，兒童能夠發現文本中涉及主角情緒、慾望、目標的因果關係。第四，隨著兒童經驗的增加，認知能力的成長，他們漸漸能夠發現文本階層結構中比較上層的關係，或是能夠辨識文本的主題、次標題。最後，則是兒童能依文本的主題推論該篇文章所要傳達的主要概念或意義是什麼。

綜合兩項閱讀發展理論，Chall（1996）的理論係依年齡階段劃分，不同年齡階段有不同的特徵與發展重點。van den Broek 等人（2005）的觀點則主要根據推論能力的發展談年齡間的差異。由於本測驗的適用對象縱跨國小二年級至高中三年級，因此擬兼採 Chall 及 van den Broek 等人的發展觀點設計命題架構。

研究一　電腦化中文閱讀能力評量之建構

DACC 除了可評估學生的整體閱讀能力以外，亦針對細項的閱讀技能進行分向度診斷，包含：字詞辨識、表層文意理解、文意統整、推論理解、分析評鑑。為使試題能有效地評估學生的閱讀能力，亦以嚴謹的流程進行試題研發，包含以下幾個階段：制訂命題結構、命題、修題、審題、預試、分析、擇題等。

（一）測驗架構

本測驗系統以 Chall（1996）的閱讀發展階段論及 van den Broek 等人（2005）所提出的理解能力發展歷程，做為命題架構的理論基礎。再者，亦參考了國際間重要的閱讀能力評量（例如，PIRLS、PISA 及 NAEP）進行測驗向度之建置。另考量本測驗系統之適用對象橫跨國小二年級至高中三年級，最終將核心的閱讀能力技能細分為五大向度：字詞辨識、表層文意理解、文意統整、推

論理解、分析評鑑。以下為各向度之定義與例題：

表 2
DACC 向度定義與例題

向度名稱	定義	例題
字詞辨識	指在閱讀的過程中，能認出文字的字面意義，或理解一個字或詞在文句裡的意思。	「人事之推移，事物之幻化，都是無常人生的景象。」中，「推移」的涵義是什麼？（A）利害 （B）改變 * （C）升遷 （D）貶謫
表層文意理解	理解文章裡所要傳達的表層意思，或是透過理解文詞涵義，擷取出需要的訊息。	「阿水的音樂細胞好像不是很多，吹口哨還可以，至於直笛和口琴，那就算了吧！」請問阿水最會吹奏什麼？（A）哨子（B）直笛（C）口哨 *（D）口琴
文意統整	統整文章中的整體訊息，瞭解文章的大意及所要傳達的主旨。	「一般人以為颶風比颱風強，其實不然。這兩者的形成方式、移動路徑與強弱是一樣的。發生在北太平洋西部及南中國海者為颱風，發生在大西洋西部、加勒比海、墨西哥灣和北太平洋東部的稱為颶風」，這段文字主要是說明颶風與颱風二者之間有何關係？（A）二者名稱不同，威力大小也不相同（B）二者威力不同，但是發生地點相同（C）二者性質相同，但是發生起點不同 *（D）二者名稱相同，但是發生地點不同
推論理解	指讀者無法直接從文章中找到答案，必須統整文章句子或段落內容，才能推論而得答案的能力。	「練字是孤獨、枯燥的，這是第一道關卡。當開始提筆學習古往今來的傑作之後，你會發現，現實與腦海中的美好幻想完全相反，例如筆不聽話、字寫得不像等諸如此類迫使人把筆收進抽屜的一萬種情況，這是再正常也不過的現象。」這段話最可能由誰說出口？ （A）學習漢字的外國學生 （B）練習水彩的歷史教授 （C）教導素描的美術老師 （D）展出作品的書法大師 *
分析評鑑	透過對文章深入的理解，評斷文章的品質，掌握文章觀點及寫作手法，客觀地評估文章適當的標題和形式；並從分析中比較出不同文章的異同、優劣。	「你有遇過價錢標籤紙撕掉後，有餘膠殘留在貨品上面，無法清理乾淨的困擾嗎？其實你只要用吹風機把標籤吹熱，便輕而易舉地將標籤紙撕下來了。」 以上這段訊息最適合刊載在下列哪一類專欄文稿中？ （A）〈心靈小棧〉 （B）〈生活小常識〉 * （C）〈省錢妙妙妙〉 （D）〈3C 產品比一比〉

（二）試題命製與修審

DACC 的命題者包含：國小教師、國中國文老師、高中國文老師、主修閱讀心理學之博士生、從事國語文研究之專業人員等。所有命題者在正式命題前，皆須研讀本研發計畫所提供之專業命題文件，並通過訓練後，方可編擬試題。命題訓練以工作坊型式進行，由計畫研究人員說明測驗目的、測驗架構、向度定義與內涵……等相關資訊後，再由長期參與 DACC 審題的教授擔任講師，講解 DACC 命題綱領，帶領命題者進行命題實務練習並給予回饋，確認命題者充分瞭解測驗內涵及具備命題知能後，再正式加入命題團隊。

在命題素材方面，以學生熟悉之題材為主，主題不限於語文，亦涵蓋生活知識、歷史、地理、及科學範疇等。文本的形式包含連續文本與非連續文本；文體有記敘文、說明文、論說文、操作指南、紀錄及超文本（hypertext）等，以呼應學生生活中可能遭遇的真實閱讀情境。題目多為題組題，題組文本由命題者自撰或以現有題材改寫，並設計各個向度的試題。題組內的每一道子題皆可對應至上述其中一個向度，藉由此設計，當學生完成測驗後，便可評估其各向度的表現。

當命題者完成試題初稿，下一階段便以會議形式，利用小組討論，相互修整題目，以釐清或調整個人命題時的盲點。為了更加確保試題的品質，所有題目皆會送交國文系教授與測驗專家進行審

查，審查重點包含：試題內涵是否符合其所預定評量之向度、正答選項之有效性、個別選項是否具備足夠的誘答力、文本與試題難度之適當性、文本與試題內容是否具公正性……等。各向度評量重點請見表 3。

表 3
DACC 各向度評量重點

向度名稱	評量重點
字詞辨識	理解字詞意義、多義字詞區辨、成語運用
表層文意理解	句意理解、擷取訊息、辨識出問題的重要元素（例如：人、事、時、地、物、方法等）
文意統整	指出文章大意、歸納文章中的多項訊息
推論理解	指稱判斷、因果推論、情節發展預測、喻意推論、推估上層概念
分析評鑑	判斷作者的寫作意圖、判斷內容邏輯是否具合理性、設定文章標題、比較文本間的關連或異同、分析文章內容之類型

（三）預試設計

試題經過命、修、審流程定稿後，本測驗亦藉由預試蒐集學生的實際作答反應，以觀察試題被學生作答的情形是否符合命題者之設計，並利用所收集到的作答反應進行後續參數估計。

1. 題本編製

DACC 為電腦化適性測驗，建立題庫時所有試題之參數須建立於共同量尺之上。由於整體試題數量頗多，每位學生僅能作答部分預試試題，故將所有題目依難度差異編製為不同題本。考量被納入題庫的試題可能來自不同預試題本，本研究在各題本間配置共同題，以利不同題本的試題難度能夠進行等化（equating），進而建立共同的難度量尺。

本測驗所有題本所欲測量的特質，皆涵蓋上述提及之五向度閱讀技能，唯受試對象因年級不同而能力有所差異，各題本的整體難度亦可能不相等，故須進行垂直等化。參考多種等化設計（Kolen & Brennan, 2004）之後，DACC 採用定錨不等組 （non-equivalent groups with anchor test, NEAT）之等化設計，如表 4 所示。在 NEAT 設計下，有部分試題同時出現在不同題本中，稱為共同題或定錨題（anchor items）。共同題數約為預試題本總題數的 20% 至 30%，每個題本中的每個向度均包含共同題，該共同題除了能連結不同題本，同時也能作為未來新發展試題加入題庫中之連結依據。表 4 以兩個測驗題本施測於兩群不等價考生為例，所有施測試題可分為 A、B 與 C 三部分，第一個題本包含試題集合 A 與 B，並施測於第一群考生；第二群考生作答的題本則包含試題集合 A 與 C。顯然，試題集合 A 為兩題本之間的共同部分，這些共同題為進行參數等化時的連結依據。

本研究利用 NEAT 等化設計，將多個題本資料合併成單一資料矩陣，並針對此矩陣進行同時估計（concurrent calibration），如此一來，所獲得的試題難度便屬於相同量尺（Kolen & Brennan, 2004），不同題本的試題難度亦可互相比較。

表 4
NEAT 等化設計

受試樣本	試題集合		
作答第 1 個題本之考生群	A	B	
作答第 2 個題本之考生群	A		C

在試題配置方面，所有預試題目皆透過專業命題者，按照文本特性與試題難度，設定初始年級難度，以作為安排受測對象之參考。例如：四年級難度之試題，其施測對象以三至五年級學生為主；五年級難度之試題，其受測對象便為四至六年級學生，以此類推。在題數安排上，由於每位學生的作答時間約為 30 至 40 分鐘，依不同難度之試題組合，各題本總題數介於 32 至 42 題。以本測驗最近一次預試為例，共施測 10 個題本、233 道題目，試題難度預估介於一至三年級之間。每個題本配有 22 至 25 題新題以及 11 題共同題。所有題本之共同題皆相同，新題在各題本間則不重複。預試題本試題配置與施測人數規劃請見表 5。以題本 F 為例，包含 11 道共同題與一至二年級難度試題 9 題、二年級難度試題 5 題、二至三年級難度試題 9 題，共 34 題。而題本 F 的施測對象為一至四年級學生各 100 人。

表 5
108 學年 DACC 預試題本規劃

預估難度		各題本題數分布									
		A	B	C	D	E	F	G	H	I	J
一年級		7	7	8	8	8					
一～二年級		6	6	6	6	7	9	9			
二年級		6	6	5	5	5	5	5			
二～三年級		4	4	4	5	4	9	5	9	8	7
三年級								6	14	15	15
共同題		11	11	11	11	11	11	11	11	11	11
	年級	人數									
施測對象	一	160	160	160	160	110	100	100			
	二	160	160	160	160	110	100	100	110		
	三					110	100	100	110	160	160
	四						100	100	110	160	160

2. 受試者取樣

除了透過上述設計進行題本編排以外，本研究亦考量受試者特性，納入城鄉差異之因素進行分層隨機抽樣，以利試題參數估計結果更能適用於母群。城鄉差異的分級標準以劉介宇等人（2006）為架構。該研究將全臺各區域鄉鎮，依人口密度、教育程度、年齡比例等變數，區分為七個都市化程度之集群，本測驗參考其分類，進一步將其整合成都市化程度高、中、低三類，並按照教育部統計處所公布之全臺高級中學、國民中學、國民小學名錄，計算此三類學校之比例，並依此比例進行隨機抽樣，以使取樣更貼近母群特性。以本測驗在 108 學年的最近一次預試為例，按照教育部統計處的學校名錄資訊與劉介宇等人之分類，108 學年全國公私立小學共計 2,633 所，分布於都市化程度高、中、低之區域比例依序為 3：4：3，班級平均人數約為 23 人，故各年級所需之施測班級數量如下表 6 所示。

表 6
108 學年 DACC 預試取樣班級數分布

年級	施測人數	預估班級數	都市化程度班級數		
			高	中	低
一	910	40	12	16	12
二	1130	50	15	20	15
三	650	29	9	12	8
四	540	24	7	10	7

3. 施測流程

所有施測之相關事宜，皆由本計畫人員與各校教務處進行聯繫與安排，施測地點為各校之電腦教室，透過網路以電腦化測驗固定題本的形式進行測驗，並由我方研究人員與各校教師共同進行監考。施測時間為一節課（約為 35 至 45 分鐘，含指導語說明時間）。在施測過程中，亦紀錄異常作答行為（如，答時間過短、無心作答等），作為後續剔除無效資料之參考。

（四）資料分析

1. 分析方法

每道試題（皆為二元計分的選擇題，答對計為 1 分，答錯計為 0 分）在預試階段皆蒐集至少 300 人以上的作答反應，並分別採用試題反應理論（item response theory, IRT）與古典測驗理論（classical test theory, CTT）進行資料分析。以下針對兩種方法個別進行說明：在 IRT 分析上，分析模式的選擇對試題參數的估計及推論具有至關重要的影響（盧宏益等人，2011），DACC 雖然在單一試題僅測量單一向度能力，但整份題本則同時測量五個向度，屬於題間多向度測驗（between-item multidimensional test）；因此，本測驗利用 Adams 等人（1997）提出的多向度隨機係數多項邏輯模式（multidimensional random coefficients multinomial logit model, MRCMLM）進行資料分析。此模式除了能估算受測者在各向度之能力參數與各試題難度之外，還能估計各向度間的相關性。相關研究顯示出向度間若具高相關性將能有效降低測量誤差、提升測驗信度並增加評估精準度性（Wang & Chen, 2004）。本測驗採用 ConQuest（Wu et al., 2007）軟體進行資料分析，以邊際最大概似估計法（marginal maximum likelihood estimation, MMLE）估計試題參數；以貝氏期望後驗分佈法（expected a posterior, EAP）進行能力參數估計。

而在 CTT 分析上，則針對題庫中每道試題計算通過率（percentage passing）、鑑別度（item discrimination），並進行選項分析，做為試題品質之觀察指標。通過率的計算方式為該題之答對人數除以作答該題之總人數，鑑別度為受測者測驗總分與答對該題與否的點二系列相關係數（point-biserial correlation）。

2. 擇題標準

為了讓 DACC 題庫能具備良好的心理計量特性，例如：各向度試題皆符合單向度假設、建立等距量尺來衡量定位所有受測者能力落點與試題難度高低等，需根據分析結果排除不符合 MRCMLM 模式之試題（misfit items），僅具備良好適配度統計量（fit statistics）的試題才能被納入題庫，以確保題庫中每道試題都具有良好的品質。本測驗以訊息加權均方適配度值（information-weighted mean-square fit statistic, infit MNSQ）為判斷試題品質的主要指標。infit MNSQ為自由度為1的卡方統計量，數值介於 0 與無限大之間，期望值為 1.0（Bond & Fox, 2015）。當某試題之 infit MNSQ 超出 0.6 至 1.4 區間之外，則視為不適配試題（Wright & Linacre, 1994），不適配試題在分析過程中將被逐步排除，直到所有試題的 infit MNSQ 皆符合小於 1.4，方結束分析過程。除了檢視前述 infit MNSQ 指標以外，本研究亦檢驗 CTT 的試題鑑別度，唯有 infit MNSQ 介於 0.6 至 1.4 之間、CTT 鑑別度達 .3 以上的試題被視為具備良好心理計量特性，方可被納入最終題庫。

（五）結果

透過上述方法進行試題分析與篩選試題後，DACC 正式題庫目前保有 1,019 道試題。表 7 為各向度試題難度的描述統計與難度分布，可發現在所有向度中，以「推論理解」的題數最多，「分析評鑑」的題數較少，後續可持續擴充此向度試題題量。在難度分布方面，可發現所有向度的難度分布皆大於正負 2.0 的範圍，已含括大部分學生的能力表現，適合用以評估目標對象（二至十二年級）的中文閱讀能力。另一方面，考量各向度的平均難度介於 -0.69 至 0.1，未來擴展題庫時，可以增加各向度難題作為命題目標，以利進一步提升高能力學生（如，高中學生）的測量精準度。

表 7
正式題庫試題描述統計

向度名稱	題數	難度分布	難度平均	難度標準差
字詞辨識	221	-4.00 ～ 3.09	-0.69	1.42
表層文意理解	178	-3.51 ～ 2.23	-0.67	1.06
文意統整	193	-3.29 ～ 2.38	-0.36	1.01
推論理解	290	-3.36 ～ 2.84	-0.42	1.15
分析評鑑	126	-2.36 ～ 2.42	0.12	1.04

經過上述題庫建置程序後，可發現 DACC 的試題具有優良的品質，可有效評量學生的閱讀理解能力。考量紙筆測驗與固定題本的測驗型式有諸多限制，為避免僵化的測驗型式影響能力評估精準度，本研究進一步規劃以電腦化適性測驗（computerized adaptive testing, CAT） 的方式建構正式施測系統，相關細節將於研究二進行說明。

研究二　正式中文閱讀能力電腦化適性測驗之建置與驗證

CAT 為現今評量的趨勢，在教育測驗相關領域的應用日趨廣泛（Keller-Margulis et al., 2018; Layes et al., 2019; Martin & Lazendic, 2018）。相較於紙筆測驗，採用 CAT 的優勢為受測者僅需接受較少量的施測題數，卻可達到與紙筆測驗相同的測量精準度（Sand et al., 1997; Wang & Chen, 2004）。進行 CAT 施測時，受測者首先接受電腦選出的測驗題目，然後依據受測者的作答反應，估計出受測者的初步能力估計值（initial ability estimate）（van der Linden & Glas, 2000）。接著，電腦根據這個初步能力估計值，從題庫挑出最適合受測者作答（例如，能最大化降低能力估計誤差）的題目讓受測者作答。這個施測過程一直持續到已經施測事先設定的題數，或是能力估計精確性已達到預先設定的要求（Thissen & Mislevy, 1990）。所以，在 CAT 裡，能力高的受測者不需作答很簡單的試題；能力低的受測者也不需作答很困難的試題，因此得以降低測驗長度，節省施測時間（Weiss, 1982）。以下針對施行 DACC 適性測驗時的初始選題、能力估計、選題策略及終止條件等重要程序進行說明。

（一）初始選題

在適性施測的前提下，試題難度必須能夠配合考生的能力水準。在學生首次施測時，由於無法確知其閱讀能力水準，故本研究根據預試結果所建立的能力常模，取學生實際年級之各向度平均能力值，並加減一固定常數（0.4），形成各向度之初始能力區間，並從區間中隨機抽取一個數值，作為該位學生的起始能力。接著，以初始能力向量計算題庫中所有題組對此能力向量提供之訊息量矩陣，並選擇訊息量矩陣之行列式值（determinant）最大的該題組作為起始試題。

（二）能力估計

當學生作答完一道題組後，DACC 即利用其作答反應以及各題的難度參數，即時估計該位學生當下的能力向量。DACC 採用貝氏最大後驗分佈法（maximum a posterior, MAP）進行能力估計，透過將五個能力向度間的變異共變數矩陣作為能力先驗分佈（Segall, 1996），可提升各向度能力估計的精準度（Wang & Chen, 2004）。MAP 法使用 Newton-Raphson 法進行疊代（iteration），持續在每次疊代過程中算出新的能力變化向量，直到各向度能力變化量之絕對值皆小於事前設定值（如，0.005），進而獲得更新後的能力估計值。相較於 EAP 法，當評量向度較多時（如，五向度），採用 MAP 法進行能力估計可縮短運算時間（Segall, 2000）。

（三）選題策略

在選題策略方面，DACC 採用最大訊息選題法（Lord, 1977），將受測者能力估計值與試題難度值代入費雪訊息函數（Fisher's information function）計算每道試題之訊息量，進而選取題庫中能提供最大訊息的試題供受測者作答。

由於 DACC 共有五個向度，本研究採用多向度適性測驗之選題算則進行選題（Frey et al., 2016），首先根據受測者能力向量與試題難度值計算每道試題所提供的訊息量矩陣，加總相同題組下所有試題的訊息矩陣後除以該題組之試題數，據此計算出每個題組所提供之平均試題訊息矩陣（Murphy et al., 2010），最終，計算每個平均試題訊息矩陣之行列式值。此外，為控制試題曝光率，本研究亦參考 Green 等人（1984）的建議，從行列式值最大高的五個題組中，隨機抽取一組作為下一個施測題組。

（四）終止條件

考慮到在學校場域中有課程安排之實務考量，僅能在有限時間（如，35 至 40 分鐘）進行施測，因此 DACC 採用固定題長作為測驗終止條件。各年級考生之施測題數分別訂為：三年級（含）以下 29 題、四至六年級 33 題、七至九年級 38 題、十年級（含）以上 44 題。

（五）常模建構

本測驗為提供使用者測驗結果的參照標準，故在電腦化適性測驗正式上線後，持續收集各年級學生的作答資料與能力值，以實徵資料為本，統計各年級學生的平均能力值，建置常模，方便使用者理解自我表現對比整體同年級學生程度的相對定位。

常模資訊收集的資料期間為自本測驗正式開放至 2019 年 6 月止，時間約為 3 年半。資料來源包含全臺所有縣市與連江縣，共有 1,255 所學校、38,099 位學生參與，其中人數佔比較高（超過 10%）的縣市依序為：新北市（15%）、臺中市（14%）、臺北市（13%）、桃園市（11%）、高雄市（11%）、臺南市（10%）。各年級人數如下表統計：

表 8
最終版本常模各年級有效受測人數

年級	一	二	三	四	五	六
人數	146	789	3,301	6,469	3,331	2,762
年級	七	八	九	十	十一	十二
人數	3,972	3,020	2,315	11,138	552	304

此外，為方便使用者判讀測驗結果，本測驗將六年級各向度之平均能力皆固定為 0，以建置固定的參照標準。但在實務上分析作答者能力值時，往往可發現，六年級的平均能力會稍有偏移，此時便會採用平移的方式統一調整所有的數值。例如：若依分析結果，「字詞辨識」的六年級平均能力值為 -0.14，便將此向度所有年級的平均能力值皆加上 0.14。此項作法不僅可使六年級的平均能力固定為 0，亦保有各年級平均能力的間距。表 9 為經平移後，各年級學生於各向度的能力值分布情形。

表 9
各年級學生常模平均能力值

年級	字詞辨識	表層文意理解	文意統整	推論理解	分析評鑑	整體閱讀能力
2	-1.64	-1.52	-1.36	-1.40	-1.24	-1.44
3	-1.24	-1.04	-1.00	-0.92	-0.84	-1.01
4	-0.76	-0.60	-0.64	-0.60	-0.48	-0.62
5	-0.36	-0.12	-0.12	-0.20	-0.12	-0.18
6	0.00	0.00	0.00	0.00	0.00	0.00
7	0.12	0.16	0.32	0.20	0.28	0.22
8	0.28	0.32	0.48	0.40	0.40	0.37
9	0.48	0.56	0.64	0.68	0.56	0.58
10	0.80	1.00	0.92	1.04	0.76	0.90
11	1.08	1.32	1.40	1.52	1.24	1.30
12	1.24	1.48	1.52	1.72	1.36	1.46

此套常模目的為提供清楚、客觀的標準，協助 DACC 的受測者在完成測驗後，評估閱讀能力的年級水準，亦幫助教師瞭解學生的閱讀能力是否符合應有的程度，以作為後續教學之參考。

結果

（一）信度檢驗

1. 重測信度

本測驗主要的用途之一為用作形成性評量，供教學現場追蹤學生的閱讀能力，故已有多所學校長期且持續性地實施 DACC。在諸多使用者中，本研究擷取重複受測的學生，以他們的測驗總成績進行重測信度的檢驗。考量多數的使用情境為在學期開始與學期末時分別進行施測，且本國的學期週數大多介於 20 至 22 週，故抓取兩次施測日期間隔 15 至 25 週之學生成績，進行相關係數檢驗。以總數 1,449 位學生的兩次測驗總分分析，可得兩者的相關係數為 .76，顯示同一位學生的兩次測驗結果具有高度相關。

2. IRT 信度

CAT 為針對每位受測者量身訂做的測驗，不同受測者接受不同的試題組合，因此每人有其個別化的測量精準度（Thissen & Wainer, 2001），稱為條件化信度（conditional reliability）（Raju et al., 2007）。本研究針對近一年（2019 年 9 月至 2020 年 10 月）16,479 位 DACC 受測者的各向度能力值，計算其條件化信度，統計結果如表 10 所示。可發現在多數向度中，平均信度值皆高於 .84、標準差僅為 0.01，顯示在這些向度上的能力估計值皆具有良好的精準度。「分析評鑑」的平均信度為最低，標準差為 0.04、相對較大，由於「分析評鑑」的試題較其他向度少，所能涵蓋的能力範圍（-2.36 ～ 2.42）亦相對較小，推測此為造成該向度信度相對較低之可能原因。

表 10
各向度條件化信度統計

向度	平均數	標準差	最小值	最大值
字詞辨識	.84	.01	.77	.87
表層文意理解	.85	.01	.76	.88
文意統整	.84	.01	.77	.87
推論理解	.84	.01	.77	.88
分析評鑑	.69	.04	.52	.77

進一步將所有受試者依據能力值高低區分成 12 個不同區間，並針對各區間所有受試者計算平均條件化信度，結果如表 11 所示。可觀察到平均信度介於 .77 ～ .82 之間，且當受測者能力值高於 -1.5 時，其信度皆超過 .80。這些結果指出，本測驗對不同閱讀水準的學生皆具有穩定良好的測驗信度。

表 11
不同閱讀程度學生之平均條件化信度統計

能力值條件範圍	人數	平均信度
小於 -2.5	104	.77
介於 -2.5 至 -2	219	.78
介於 -2 至 -1.5	354	.79
介於 -1.5 至 -1	1,069	.80
介於 -1 至 -0.5	2,089	.81
介於 -0.5 至 -0	2,634	.81
介於 0 至 0.5	2,409	.82
介於 0.5 至 1	2,708	.82
介於 1 至 1.5	2,401	.82
介於 1.5 至 2	1758	.82
介於 2 至 2.5	602	.81
大於 2.5	132	.80

（二）效度檢驗

CAT 的效度可分別從試題層次與測驗層次進行驗證，在試題層次，主要檢視是否測驗是否具備單向度性（unidimensionality），意即是否所有試題皆評量相同的潛在特質。一旦測驗符合單向度假設，即便不同受測者作答測驗中不同試題（如，CAT 之施測情境），亦能確保測量結果之有效性（Wainer et al., 2000）。在測驗層次，主要檢視 CAT 測驗結果與另一個測量相似建構的測驗（如，評量閱讀能力的紙本測驗），兩者之間是否有關連性存在。

1. 效標關聯效度

為驗證 DACC 的測驗結果可有效反映學生的閱讀能力，本研究以大型標準化測驗為效標，抓取 2019 年接受 DACC 測驗的 2,332 位九年級學生資訊，向國立臺灣師範大學心理與教育測驗研究發展中心申請該批學生的國中教育會考國文科測驗結果。國中教育會考為本國參與人數最多之標準化測驗，每年約有 21 萬名國中畢業生應試。其中，該測驗國文科試題主要評量的能力包含語文知識、文意理解、文本評鑑，與 DACC 所評量的能力雖不完全相同，但著重閱讀能力的核心評量價值相似。

以 2019 年會考國文科試題為例，整份試卷共有 48 道試題，評量語文知識、文意理解、文本評鑑的題數分別為 11 題、32 題、5 題。利用學生的 DACC 整體閱讀能力成績與 2019 年會考國文科試題答對題數進行效標關聯效度之檢驗，結果可發現兩者的相關係數為 .64，為中度相關，這代表當受測者在 DACC 的整體閱讀能力成績偏低（或較高）時，其參加會考國文科測驗的答對題數可能也偏低（或較高），因此 DACC 的測驗成績可有效的評估學生的閱讀能力。

2. 建構效度

為驗證 DACC 各向度試題是否具備單向度性，本研究針對各向度試題之預試資料進行分析，進而算出每個試題的 infit MNSQ，作為試題層次的效度驗證指標（Baghaei, 2008）。當某試題之 infit MNSQ 超出 0.6 至 1.4 區間之外，則視為不適配試題（Bond & Fox, 2015; Wright & Linacre, 1994），不適配試題在分析過程中將被逐步排除，直到所有試題的 infit MNSQ 皆符合小於 1.4。經過分析後，本測驗有將近 85% 的預試題目都能符合 MRCMLM 模式。經過刪除不適配試題後，目前題庫中各向度全數試題皆符合本測驗所採用的分析模式，故可知 DACC 各向度試題具備良好的建構效度。

討論與建議

有鑑於教學現場於閱讀能力評量工具的匱乏，本研究旨在開發有效、且能長期應用於課堂的中文閱讀能力診斷測驗。現有閱讀測驗多存在僅適用部分年級、僅能單次施測、無法診斷細項閱讀技能程度、或不提供個人成績之限制，因此學校教師往往無法明確掌握學生的閱讀能力發展情形，而 DACC 則是解決此困境的有效方案。

現有的中文閱讀測驗，如，國小學童中文閱讀理解測驗（王木榮、董宜俐，2006）、閱讀理解量表（張世彗，2014），測驗目標對象僅包含單一年級或兩個年級，致使測驗僅能有一次性的應用，難以廣泛地應用於教學現場。而 DACC 有效地解決了此問題，將測驗的適用範圍擴增為二至十二年級，使學生於重要的閱讀能力發展時期能全程瞭解自我閱讀理解能力的成長情形。

同時，為了擴展適用對象，DACC 在命題時便依不同年級題材進行命題，使測驗文本的取材與難度可以更貼合不同年齡、不同程度的學生作答；為了能持續追蹤學生的能力變化情形，DACC 利用現代測驗理論分析試題難度，透過等化的程序將所有題目的難度參數建構於同一量尺上，因此縱使學生作答的題目不同，依其答對情況所估計出來的能力值亦能相互比較，使不同學生的測驗成績能夠互比以外，單一學生多次受測的結果亦能持續記錄與比較。此項優勢補足了市面上多數閱讀理解測驗僅有單一版本，如，國小句型理解測驗（張祐瑄、蘇宜芬，2014）、中文閱讀理解測驗（林寶貴、錡寶香，2000）等；或雖有複本但跨年級題本間因無等化而使分數無法互比（如，閱讀理解成長測驗）（蘇宜芬等人，2018）的弱點。

而在診斷閱讀細項技能方面，DACC 為題間多向度，每道題目可對應至字詞辨識、表層文意理解、文意統整、推論理解、分析評鑑其中一個向度。在回答完整份測驗後，便可依各向度試題的答題情形，評估出學生在這些向度的表現，瞭解不同向度間是否存在能力落差，可做為調整教學重點的依據。除此之外，DACC 呼應國際測驗趨勢（如，TOEFL、GRE 等），以電腦化適性測驗的形式建構測驗系統，不僅縮短施測時間，且可於施測完後即時提供測驗結果報表，並透過教師權限的設計，使學校教師掌握每一位學生的閱讀能力表現。當班內學生的程度不一時，相關成績資訊便是提供差異化教學最有效的參考依據。

DACC 除了在上述層面解決了現有閱讀能力評量工具的限制以外，本研究亦驗證了 DACC 的測驗信度與效度。以重測信度而言，藉由 1,449 位重複受測兩次、間隔約四至六個月期間的成績資料檢視，其兩次測驗的相關係數高達 .76，顯示 DACC 的評測結果具有良好的穩定性。而在效標關聯效度方面，本研究亦驗證了 DACC 與大型標準化相關領域的測驗具有中度相關，且透過嚴謹的 IRT 試題分析管控，所有試題皆具有優良的品質，使 DACC 可有效評量出學生的閱讀能力表現。這些特點為 DACC 具備現有評量工具所無法超越的優勢，提供了實徵證據。

而在研究限制方面，由於 DACC 題庫試題多為題組形式，題組中各試題的難度難免有所差異，在 CAT 施測過程下，被選出的最佳題組中，可能有少數試題的難度與能力參數之差異相對較大，

致使雖然該題所提供的訊息量較低，但因該題屬於此最佳題組而仍被施測。在測驗初期，由於能力參數估計誤差較大，對於能力之可能落點尚不明確，若題組中各試題難度較為分散，應不構成太大影響；但在測驗後期能力參數估計誤差較小時，施測訊息量較低的試題，將拉低測驗的施測效能，此為本系統的限制，未來宜針對此點進行改善。

最後，在未來研發上，DACC 將持續擴展題庫規模，除了加入更多單題以外，亦可強化難度較高的試題，並增加分析評鑑向度的試題數量，為日益增長的受測人數提前籌備。此外，目前 DACC 並未配有試題曝光度控制的機制，在未來題庫數量達到有效地擴展後，可逐步加進更多試題管控的機制，使測驗派題更臻完美。在測驗效度上，亦可再擴增驗證範圍，例如：收集大學生的受測資料，並針對主修中文系和非中文系學生的成績進行分析，比較兩者之間的閱讀理解能力是否有明顯差異，作為區辨效度的證據。另一方面，在閱讀能力發展的相關領域上，仍有許多議題尚待探討，例如，基礎閱讀能力與高層次閱讀理解技能之間是以序列式進行發展，抑或為同時、各自獨立發展，現有文獻仍存有歧異（如，陳茹玲、宋曜廷等人，2017；Rapp et al., 2007; Schmitt et al., 2011），DACC 作為跨學習階段的閱讀理解能力評量工具，便可為此提供實徵證據進行驗證，對閱讀理解領域做出更多貢獻。

參考文獻

王木榮、董宜俐（2006）：《國小學童中文閱讀理解測驗》。心理出版社。[Wang, M.-R., & Dong, Y.-L. (2006). *Reading comprehension test for elementary school students.* Psychological Publishing.]

沈欣怡、蘇宜芬（2011）：〈推論性問題引導課程對國小四年級學童推論理解與閱讀理解能力之影響〉。《教育心理學報》，*43*（S），337–356。 [Shen, H.-Y., & Su, Y.-F. (2011). The effects “inferential question discussion program” on inferential comprehension and reading comprehension of fourth grade students. *Bulletin of Educational Psychology, 43*(S), 337–356.] https://doi.org/10.6251/BEP.20110801

林小慧、曾玉村（2017）：〈科學多重文本閱讀理解評量之建構與信效度分析—以氣候變遷與三峽大壩之間的關係題本為例〉。《教育心理學報》，*49*（2），215–241。[Lin, H.-H., & Tzeng, Y.-H. (2017). Developing and validating a scientific multi-text reading comprehension assessment: Evidence from texts describing relationships between climate changes and the Three Gorges Dam. *Bulletin of Educational Psychology, 49*(2), 215–241.] https://doi.org/10.6251/BEP.2017-49(2).0003

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 the test of language comprehension. *Bulletin of Special Education, 19*, 79–104.]

柯華葳（1999）：〈閱讀理解困難篩選測驗〉。《測驗年刊》，*46*（2），1–11。[Ko, H.-W. (1999). Reading comprehension screening test. *Psychological Testing, 46*(2), 1–11.]

柯華葳、張郁雯、詹益綾、丘嘉慧（計畫主持人）（2017）：《PIRLS 2016 臺灣四年級學生閱讀素養國家報告》（計畫編號：MOST 102-2511-S-008-018-MY4）。教育部委託專案報告，國立中央大學。https://drive.google.com/file/d/1FwGC4tNZ7O8y9c1J3BGTovgDcdQ8FFZK/view [Ko, H.-W., Chan, Y.-L., Chang, Y.-W., & Chiu, J.-H. (Principal Investigator). (2017). *PIRLS 2016 report for fourth-grade: Students reading literacy in Taiwan* (Report No. MOST 102-2511-S-008-018-MY4) (Grant). National Central University. https://drive.google.com/file/d/1FwGC4tNZ7O8y9c1J3

BGTovgDcdQ8FFZK/view]

柯華葳、詹益綾（2006）：《國民小學（二至六年級）閱讀理解篩選測驗》。教育部特殊教育小組。[Ko, H.-W., & Chan, Y.-L. (2006). *Reading comprehension screening test for second to sixth graders.* Department of Student Affairs and Special Education, Ministry of Education.]

張世彗（2014）：〈閱讀理解量表建製之探究〉。《特殊教育發展期刊》，*58*，1–12。[Chang, S.-H. (2014). Test making of reading comprehension. *The Development of Special Education, 58*, 1–12.] https://doi.org/10.7034/DSE.201412_(58).0001

張祐瑄、蘇宜芬（2014）：〈「國小句型理解測驗」之編製及其信、效度研究報告〉。《測驗學刊》，*61*（3），385–408。[Chang, Y.-H., & Su, Y.-F. (2014). The reliability and validity of sentence comprehension test for elementary school students. *Psychological Testing, 61*(3), 385–408.]

陳昭珍、宋曜廷、章瓊方、曾厚強（2020）：〈配合國小課程單元科普讀物人工分級推薦與系統可讀性分析之差異研究〉。《圖書資訊學刊》，*18*（1），45–67。[Chen, C.-C., Sung, Y.-T., Chang, C.-F., & Tseng, H.-C. (2020). Examining the differences of readability leveling of Chinese popular science books by experts and by CRIE system for elementary school children. *Journal of Library and Information Studies, 18*(1), 45–67.] https://doi.org/10.6182/jlis.202006_18(1).045

陳茹玲、宋曜廷、蘇宜芬（2017）：〈「精緻化推論教學課程」對國小弱勢低年級學生策略運用、閱讀理解與故事重述表現之影響〉。《教育心理學報》，*48*（3），303–327。[Chen, J.-L., Sung, Y.-T., & Su, Y.-F. (2017). The effect of "elaboration curriculum" on the reading strategy, reading comprehension and story retelling of 2nd grade students. *Bulletin of Educational Psychology, 48*(3), 303–327.] https://doi.org/10.6251/BEP.20150922

陳茹玲、曾厚強、宋曜廷、林慶隆、柯華葳（2017）：〈華語文教材之文本分析與可讀性研究〉。《國際中文教育學報》，*1*，39–71。[Chen, J.-L., Tseng, H.-C., Sung, Y.-T., Lin, C.-L., & Ko, H.-W. (2017). Analyzing the readability of text for Chinese as foreign language learners. *International Journal of Chinese Language Education, 1*, 39–71.]

劉介宇、洪永泰、莊義利、陳怡如、翁文舜、劉季鑫、梁賡義（2006）：〈臺灣地區鄉鎮市區發展類型應用於大型健康調查抽樣設計之研究〉。《健康管理學刊》，*4*（1），1–22。[Liu, C.-Y., Hung, Y.-T., Chuang, Y.-L., Chen, Y.-J., Weng, W.-S., Liu, J.-S., & Liang, K.-Y. (2006). Incorporating development stratification of Taiwan townships into sampling design of large scale health interview survey. *Journal of Health Management, 4*(1), 1–22.] https://doi.org/10.29805/JHM.200606.0001

盧宏益、徐永豐、薛國松（2011）：〈模式錯誤假設對電腦化測驗的影響〉。《教育心理學報》，*42*（4），613–630。[Lu, H.-Y., Hsu, Y.-F., & Hsueh, K.-S. (2011). The effect of model misspecification on computerized testing. *Bulletin of Educational Psychology*, *42*(4), 613–630.] https://doi.org/10.6251/BEP.20100302

謝進昌（2019）：〈促進中文閱讀理解教學成效量化研究統合：調節變項影響與評估〉。《教育科學研究期刊》，*64*（4），175–206。[Hsieh, J.-C. (2019). Synthesis of quantitative research on Chinese reading comprehension instruction: Analysis of moderating effects. *Journal of Research in*

Education Sciences, 64(4)*,* 175–206.] https://doi.org/10.6209/JORIES.201912_64(4).0007

蘇宜芬、洪儷瑜、陳柏熹、陳心怡（2018）：〈閱讀理解成長測驗之編製研究〉。《教育心理學報》，*49*（4），557–580。[Su, Y.-F., Hung, L.-Y., Chen, P.-H., & Chen, H.-Y. (2018). The development of progress monitoring test of reading comprehension. *Bulletin of Educational Psychology, 49*(4), 557–580.] https://doi.org/10.6251/BEP.201806_49(4).0003

Adams, R. J., Wilson, M. R., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement*, *21*(1), 1–23. https://doi.org/10.1177/0146621697211001

Baghaei, P. (2008). The Rasch model as a construct validation tool. *Rasch Measurement Transaction*s, *22*(1), 1145–1146.

Bond, T. G., & Fox, C. M. (2015). *Applying the Rasch model: Fundamental measurement in the human sciences* (3rd ed.). Routledge. https://doi.org/10.4324/9781315814698

Chall, J. S. (1996). *Stages of reading development* (2nd ed.). Harcourt Brace.

Crossley, S. A., & McNamara, D. S. (Eds.). (2016). *Adaptive educational technologies for literacy instruction.* Routledge. https://doi.org/10.4324/9781315647500-1

Frey, A., Seitz, N.-N., & Brandt, S. (2016). Testlet- based multidimensional adaptive testing. *Frontiers in Psychology*, *7*, Article 1758. https://doi.org/10.3389/fpsyg.2016.01758

Green, B. F., Bock, R. D., Humphreys, L. G., Linn, R. L., & Reckase, M. D. (1984). Technical guidelines for assessing computerized adaptive testing. *Journal of Educational Measurement, 21*(4), 347–360. https://doi.org/10.1111/j.1745-3984.1984.tb01039.x

Hong, J.-F., Peng, C.-Y., Tseng, H.-C., & Sung, Y.-T. (2020). Linguistic feature analysis of CEFR labeling reliability and validity in language textbooks. *Journal of Technology and Chinese Language Teaching, 11*(1), 57–83.

Hsiung, H.-Y., Chang, Y.-L., Chen, H.-C., & Sung, Y.-T. (2017). Effect of stroke-order learning and handwriting exercises on recognizing and writing Chinese characters by Chinese as a foreign language learners. *Computers in Human Behavior, 74*, 303–310. https://doi.org/10.1016/j.chb.2017.04.022

Keller-Margulis, M., McQuillin, S. D., Castañeda, J. J., Ochs, S., & Jones, J. H. (2018). Identifying students at risk: An examination of computer-adaptive measures and latent class growth analysis. *Journal of Applied School Psychology, 34*(1), 18–35. https://doi.org/10.1080/15377903.2017.1328627

Kolen, M. J., & Brennan, R. L. (2004). T*est equating, scaling, and linking: Methods and practices*. Springer Publishing Company.

Lan, Y.-J., Chen, N.-S., & Sung, Y.-T. (2017). Guest editorial: Learning analytics in technology enhanced language learning. *Journal of Educational Technology & Society, 20*(2), 158–160.

Layes, S., Lalonde, R., & Rebai, M. (2019). Effects of an adaptive phonological training program on reading and phonological processing skills in Arabic-speaking children with dyslexia. *Reading and Writing Quarterly, 35*(2), 103–117. https://doi.org/10.1080/10573569.2018.1515049

Liao, C.-N., Chang, K.-E., Huang, Y.-C., & Sung, Y.-T. (2020). Electronic storybook design, kindergartners'

visual attention, and print awareness: An eye-tracking investigation. *Computers & Education, 144*, Article 103703. https://doi.org/10.1016/j.compedu.2019.103703

Lord, F. M. (1977). A broad-range tailored test of verbal ability. *Applied Psychological Measurement*, *1*(1), 95–100. https://doi.org/10.1177/014662167700100115

Martin, A. J., & Lazendic, G. (2018). Computer-adaptive testing: Implications for Students' achievement, motivation, engagement, and subjective test experience. *Journal of Educational Psychology, 110*(1), 27–45. https://doi.org/10.1037/edu0000205

Murphy, D. L., Dodd, B. G., & Vaughn, B. K. (2010). A comparison of item selection techniques for testlets. *Applied Psychological Measurement*, *34*(6), 424–437. https://doi.org/10.1177/0146621609349804

National Center for Education Statistics. (2009). *The nation's report card: Reading 2009* (NCES 2010-458). Institute of Education Sciences, U.S. Department of Education. https://nces.ed.gov/nationsreportcard/pdf/main2009/2010458.pdf

National Reading Panel. (2000). *Teaching children to read: An evidence-based assessment of the scientific research literature on reading and its implications for reading instruction* (NIH Publication No. 00-4769). National Institutes of Child Health and Human Development. https://www.nichd.nih.gov/sites/default/files/publications/pubs/nrp/Documents/report.pdf

Organization for Economic Co-operation and Development. (2010). *PISA 2009 results: What students know and can do – Student performance in reading, mathematics and science (Volume I)*. https://doi.org/10.1787/9789264091450-en

Raju, N. S., Price, L. R., Oshima, T. C., & Nering, M. L. (2007). Standardized conditional SEM: A case for conditional reliability. *Applied Psychological Measurement*, *31*(3), 169–180. https://doi.org/10.1177/0146621606291569

Rapp, D. N., van den Broek, P., McMaster, K. L., Kendeou, P., & Espin, C. A. (2007). Higher-order comprehension processes in struggling readers: A perspective for research and intervention. *Scientific Studies of Reading, 11*(4), 289–312. https://doi.org/10.1080/10888430701530417

Rosita Cecilia, M., Vittorini, P., & di Orio, F. (2016). An adaptive learning system for developing and improving reading comprehension skills. *Journal of Education Research, 10*(4), 195–236.

Sand, W. A., Water, B. K., & McBride, J. R. (Eds.) (1997). C*omputerized adaptive testing: From inquiry to operation.* American Psychological Association. https://doi.org/10.1037/10244-000

Schmitt, N., Jiang, X., & Grabe, W. (2011). The percentage of words known in a text and reading comprehension. *The Modern Language Journal, 95*(1), 26–43. https://doi.org/10.1111/j.1540-4781.2011.01146.x

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika*, *61*, 331–354. https://doi.org/10.1007/BF02294343

Segall, D. O. (2000). Principles of multidimensional adaptive testing. In W. J. van der Linden & G. A. W. Glas (Eds.), *Computerized adaptive testing: Theory and practice* (pp. 53–73). Springer Publishing Company. https://doi.org/10.1007/0-306-47531-6_3

Sung, Y.-T., Lee, H.-L., & Yang, J.-M. (2019). The quality of experimental designs in mobile learning research: A systemic review and self-improvement tool. *Educational Research Review, 28,* Article 100279. https://doi.org/10.1016/j.edurev.2019.05.001

Thissen, D., & Mislevy, R. J. (1990). Testing algorithms. In H. Wainer (Ed.), *Computerized adaptive testing: A primer* (pp. 103–134). Lawrence Erlbaum Associates Publishers.

Thissen, D., & Wainer, H. (Eds.). (2001). *Test scoring*. Routledge. https://doi.org/10.4324/9781410604729

Tseng, H.-C., Chen, B., Chang, T.-H., & Sung, Y.-T. (2019). Integrating LSA-based hierarchical conceptual space and machine learning methods for leveling the readability of domain-specific texts. *Natural Language Engineering, 25*(3), 331–361. https://doi.org/10.1017/S1351324919000093

van den Broek, P., Kendeou, P., Kremer, K., Lynch, J. S., Butler, J., White, M. J., & Lorch, E. P. (2005). Assessment of comprehension abilities in young children. In S. Stahl & S. Paris (Eds.), *Children's reading comprehension and assessment* (pp. 107–130). Lawrence Erlbaum Associates Publishers.

van der Linden, W. J., & Glas, C. A. W. (2000). *Computer adaptive testing: Theory and practice.* Kluwer Academic Publishers.

Wainer, H., Dorans, N. J., Eignor, D. R., Flaugher, R. L., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer* (2nd ed.). Lawrence Erlbaum Associates Publishers. https://doi.org/10.4324/9781410605931

Wang, W.-C., & Chen, P.-H. (2004). Implementation and measurement efficiency of multidimensional computerized adaptive testing. *Applied Psychological Measurement*, *28*(5), 295–316. https://doi.org/10.1177/0146621604265938

Weiss, D. J. (1982). Improving measurement quality and efficiency with adaptive testing. *Applied Psychological Measurement, 6*(4), 473–285. https://doi.org/10.1177/014662168200600408

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

Wu, M. L., Adams, R. J., Wilson, M. R., & Haldane, S. A. (2007). *ACER ConQuest Version 2.0: Generalised item response modeling software* [Computer software and manual]. ACER Press.

Yang, J.-M., Sung, Y.-T., & Chang, K.-E. (2020). Use of meta-analysis to uncover the critical issues of mobile inquiry-based learning. *Journal of Educational Computing Research, 58*(4), 715–746. https://doi.org/10.1177/0735633119879366

收稿日期：2020 年 11 月 25 日
一稿修訂日期：2020 年 12 月 14 日
二稿修訂日期：2020 年 12 月 22 日
接受刊登日期：2020 年 12 月 24 日

Bulletin of Educational Psychology, 2021, 53(2), 285–306
National Taiwan Normal University, Taipei, Taiwan, R. O. C.

Development and Validation of a Diagnostic Assessment of Chinese Competence System

Yi-Hsuan Lee
Research Center for Psychological and Educational Testing, National Taiwan Normal University

Yeh-Tai Chou
Department of Educational Psychology and Counseling, National Taiwan Normal University

Reading comprehension is essential for learning in all subjects and for lifelong learning; it is also a crucial ability allowing people to communicate and interact with one another. Therefore, large-scale international assessments such as the Progress in International Reading Literacy Study and Program for International Student Assessment incorporate reading as an indicator of learning outcomes. This study also recognizes the essential nature of reading comprehension.

However, existing reading comprehension tests have several limitations. For example, the target populations of most tests comprise students in specific grades (e.g., elementary school students) or groups (e.g., students with special needs), and the assessments involve paper-and-pencil tests with fixed items that requires a lot of resources on test implementation and scoring. Currently, no Chinese reading comprehension assessment suitable for long-term implementation in general classrooms exists. Accordingly, the purpose of this study was to develop an assessment system, namely the Diagnostic Assessment of Chinese Competence (DACC), for comprehensively evaluating students' reading abilities in the form of a computerized adaptive test. The reliability and validity of this system were also verified.

The DACC system holistically assesses students' reading comprehension and assesses student performance in reading subskills such as comprehension (e.g., lexical, literal, and inferential), contextual integration, and analysis and evaluation. This assessment system was designed for students from the 2nd grade to the 12th grade.

The DACC test items were drafted by school teachers, doctoral students in psychology, and professionals engaged in research on the Chinese language. All drafters were required to attend and pass training before contributing test items to the DACC system. Item topics were selected to be familiar to students, such as topics relating to daily or school life. The topics are not limited to the language arts, covering life experience, history, geography, and science. In the proposed system, assessment texts appear in various formats, including continuous texts, noncontinuous texts, mixed texts, multiple texts, and texts displayed in hypertext. Text styles are also varied and include texts written in narrative, expository, descriptive, and argumentative styles. This wide range of texts reflects real-world reading situations encountered by students in their lives. Most of the DACC items are testlets, with each of the questions in the testlet corresponding to one of the five dimensions including vocabulary, literal comprehension, contextual integration, inferential comprehension, and analysis and evaluation. Such a design measures student performance in each of the dimensions and results in a comprehensive analysis of their reading abilities upon completion of the DACC.

All test items were subjected to pilot tests to collect actual responses from students for the purpose of observing whether the questions meet the proposed design. The responses were also used to estimate item parameters. All DACC items were vertically equated on the basis of the nonequivalent groups with anchor test design. In the pilot tests, the characteristics of the respondents were also considered. Stratified random sampling was adopted to recruit students from both urban and rural areas to ensure that

the parameter estimation results for the items apply to all students in the population.

The DACC items were dichotomously scored in the pilot tests. At least 300 responses were gathered for each test item, and both classical test theory (CTT) and item response theory (IRT) were applied to analyze the responses. In the IRT-based analysis, this study used the multidimensional random coefficients multinomial logit model (MRCMLM) with marginal maximum likelihood estimation to estimate item parameters and used expected a posteriori measures to estimate ability parameters. In the CTT-based analysis, the pass rates and item discrimination were calculated for each item.

To screen the DACC items for favorable psychometric characteristics, this study adopted two indicators. In the IRT-based analysis, the information-weighted mean square fit statistic (infit MNSQ) was used as the indicator to rule out misfit items, and items with infit MNSQ values between 0.6 and 1.4 were retained. In the CTT-based analysis, item discrimination was used as the indicator. Test items with discrimination of .3 or higher were retained. Accordingly, only when test items that met the requirements for both of these two indicators were entered into the formal item bank of the DACC system, resulting in 1019 items in this bank after data analysis. The range of item difficulties are bigger than -2 to 2, which corresponds to the ability parameters that include most students. The screening also demonstrated that the DACC is suitable for assessing the reading comprehension skills of students from the 2nd to 12th grades.

To strengthen the effectiveness of the DACC system, this study constructed an assessment system based on computerized adaptive testing. For estimation of abilities, maximum a posteriori estimation (MAP) was used. For test item selection, Fisher's information was applied to calculate the item information each time students finished answering a set of questions. The system then randomly assigned the next question from the five items with the highest information score. When the number of items answered met a previously set standard, the assessment was terminated.

Furthermore, this study provided a set of reference norms for the students' test results. A total of 38,099 students from 1,255 schools in Taiwan were included in the study. For these students, average scores were calculated for the students in each grade through the DACC system. Thus, students completing the assessment could compare their results against the norm and understand the level of their performance on the test. Such a reference can provide clear and objective standards to assist DACC users in assessing the grade level of their reading abilities. Accordingly, teachers can both determine whether their students' reading abilities meet the required level and adjust their follow-up instruction based on the assessment results.

In addition to the rigorous procedures for constructing the DACC assessment system, this study examined the reliability and validity of the system. For the test-retest reliability assessment, this study evaluated the scores of 1,449 students who completed the test twice; the evaluation results revealed that the average correlation of their two scores was .76, meaning that the DACC system has high reliability. In the IRT analysis, the conditional reliability of the DACC system was also high. Assessing the test results of 16,479 students revealed that the average reliability of the system was above .80, indicating that the DACC system has a stable and high reliability level for students of differing reading abilities. The validity of the assessment system was examined on the basis of criterion-related validity. Assessing the scores of 2,332 ninth-grade students who underwent both the DACC and the Comprehensive Assessment Program for Junior High School Students (a large-scale standardized test that all graduates of junior high school in Taiwan must complete) indicated that the correlation of the scores from the two tests was moderate (.64) . Moreover, construct validity assessment results demonstrated that all DACC items fit the MRCMLM.

In summary, this study adopted a series of rigorous procedures to construct a DACC assessment system; the reliability and validity of the DACC were also verified. IRT was utilized to analyze item parameters to determine difficulty levels and student ability levels. Additionally, an item bank and ability norms were established for the system, thus enabling the use of a computerized adaptive test for assessment, which can effectively determine reading comprehension levels and provide long-term tracking of reading ability growth trends. Results of test-retest reliability, conditional reliability, criterion validity, and IRT validity tests indicate that the DACC system provides a stable and effective assessment of student reading ability. For future studies, the DACC system's item bank will be expanded. A control mechanism for the item exposure rate can also be adopted to improve the system's effectiveness. Moreover, as a comprehensive assessment tool across multiple learning stages, the DACC system can provide empirical evidence for use in solving problems related to reading comprehension and make substantial contributions to related fields of research.

Keywords: reading comprehension, computerized adaptive testing, diagnosis, Chinese reading ability