

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*

林小慧

國立台灣師範大學
科學教育研究所

曾玉村

國立中正大學
師資培育中心
暨教育學研究所

本研究主要目的係在發展「科學多重文本閱讀理解評量」，並建立一組評鑑閱讀理解能力之「多重文本閱讀理解評量標準」。本評量之科學題本為「氣候變遷與中國長江三峽大壩的關係」，包含「提取訊息」、「概化訊息」、「解釋訊息」以及「整合訊息」四個分評量，共計 10 題選擇題及 9 題建構題。分析結果顯示，評分者內之 *Cronbach's α* 值均大於 .9，表示評分者內一致性尚稱良好。其次，評分者間之 *Kendall ω* 和諧係數值大於 .8，*P* 值小於 .001，達顯著相關，顯示評分者間有相同相對等級的評分趨勢。另評分者嚴苛度之多面向 Rasch 測量模式與評定量尺及部分給分模式比較之卡方考驗則達顯著水準，表示評分者間的嚴苛度及閾值嚴苛度存在差異存在，前者 *infit* 與 *outfit MFRM* 均介於 1±0.3 之間，表示無論是嚴格或寬鬆的評分者，均能有效區分出高、低能力的學生；後者意謂著對於牽涉到評分者之詮釋、評估、評分的心理歷程，本來就很難像機器評分一樣的一致性，亦符合一般對於人評分的預期，並可被理解與接受。其次，題本之內部一致性，除「提取訊息」、「概化訊息」外，其餘亦均大於 .70，全評量 α 則在 .90 以上，顯示 SMTRCA 之 *Cronbach's α* 尚在可接受範圍內。最後，驗證性因素分析也支持「科學多重文本閱讀理解評量」四因素之假設模式，兩者適配尚稱符合。本研究初步發現「科學多重文本閱讀理解評量」可分為「提取訊息」、「概化訊息」、「解釋訊息」以及「整合訊息」四個分評量，而該四個分評量分數所表徵之一階潛在因素，可被「科學多重文本閱讀理解評量」解釋的變異量分別為 .60、.66、.80、.80。

關鍵詞：多面向 Rasch 測量模式、科學多重文本閱讀理解評量、部分給分模式、評定量尺模式、驗證性因素分析

* 本篇論文通訊作者：曾玉村，通訊方式：tccytt@ccu.edu.tw。

有鑑於現今學生獲取科學知識途徑的多元化，科學多重文本的閱讀成為日常無可避免的例行事項，亦凸顯科學多重文本閱讀理解能力的重要性。針對文本閱讀理解評量，國外學者主要測量讀者對於單一文本內容的記憶，以及較深層、情境式的理解（Royer, Carlo, Dufresne, & Mestre, 1996），也有研究還進一步測量跨本文間的推論能力（Bråten & Strømsø, 2010; Bråten, Strømsø, & Britt, 2009）。至於國內學者則有針對多重文本課外閱讀對學生歷史理解的研究（唐淑華、蔡孟寧、林烘煜，2015），以及探討閱讀目的對成人閱讀多重文本的理解影響（楊蕙如、陸怡琮，2010）。

綜上所述，國內尚無針對科學議題發展一套用以評量學生多重文本閱讀理解的測驗，即使國外已發展出文本間推論判斷作業（InterVT），藉以評測跨本文間的推論能力，亦還未有研究試圖發展評鑑多重文本閱讀理解的評量規準（criteria），藉以提供標準參照模式暨有診斷功能之分析式的回饋。此外，多重文本閱讀理解歷程的研究，只是近二十年的事。Bråten 等人指出闡釋多重文本理解的理論，從 Hartman 在文學理論中所提出的互文性概念（intertextuality），Spiro 提出的認知變通理論（cognitive flexibility theory）、到現今影響最鉅的文件模式（documents model），均仍各持一說、眾家紛紜（Bråten et al.）。因此，本研究提出影響科學多重文本閱讀理解的重要構念，並蒐集實徵資料進行測量模式的驗證，藉以建立科學多重文本閱讀理解評量的建構效度。基此，本研究為有效評測學生多重文本閱讀理解的能力，研擬發展檢具選擇反應型及建構反應型之「科學多重文本閱讀理解評量」（scientific multiple text reading comprehension assessment，以下簡稱 SMTRCA），藉以評測學生多重文本閱讀理解的能力。由於學生針對建構題作答時，必須依據題幹反應帶有文字敘述的說明。這類題型不僅可以測量敘述、情境和程序等三類知識，亦能檢測記憶能力、組織能力或想法啟發暨高層次認知能力（Valenti, Neri, & Cucchiarelli, 2003）。

Bennett 與 Ward（1993）認為建構反應評量（constructed-response assessments）相較於選擇反應評量（selected response assessments）雖檢具很多優勢，但也存在評量客觀暨耗時費力的雙重疑慮問題，亦即知覺主觀的評分歷程往往導致評分回饋的變異，致使評分的客觀性受到質疑。因此本研究第二個目的在發展多重文本閱讀理解評量規準（the Rubric of Multi-Text Reading Comprehension Assessment，以下簡稱 RMTRCA），藉以提供評分者一個統一的評分規準，裨益客觀公正地評測學生的閱讀理解狀況暨提供調整教學和閱讀策略的回饋資訊。綜上所述，本研究目的係在發展 SMTRCA 暨 RMTRCA，藉以客觀評測學生科學多重文本的閱讀理解狀況，包括字詞解讀與句型解析、內容歸納與重點概化、隱含意義的推論與文本訊息的表述，暨依據文本訊息產生觀點並據以進行文本分析。

本研究首先探討多重文本閱讀理解歷程的本質內涵，做為界定 SMTRCA 發展之構念依據；其次界定 SMTRCA 架構，以作為本研究發展 SMTRCA 的理論基礎；最後簡介多面向 Rasch 測量模式的理論應用。

一、多重文本閱讀理解歷程的本質意涵

（一）單一文本閱讀理解歷程

Kintsch（1988）提出建構－整合模式（Construction-Integration Model，簡稱 CI Model），用以闡釋單一文本閱讀理解時的認知處理模式，包括「建構」與「整合」兩個認知歷程，暨文本模式及情境模式兩個心理表徵（參照圖 1），茲闡釋如下：

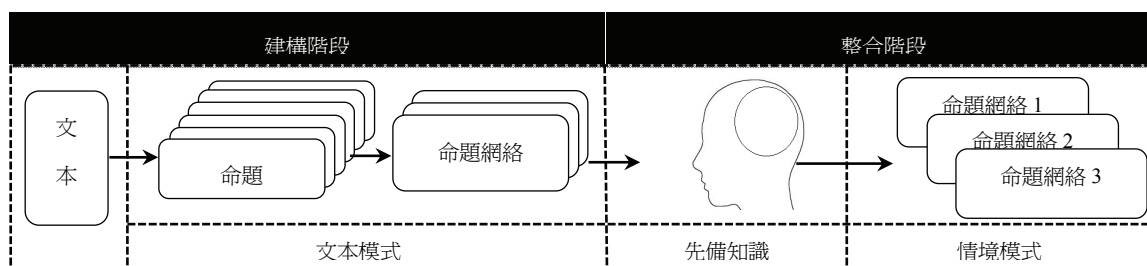


圖 1 單一文本閱讀理解歷程

1. 建構階段（construction stage）

建構階段，文本會激發讀者長期記憶中與文本相關的知識以進行解讀。讀者會藉由解析字詞的文法及意義，激發對字詞的理解知識，轉換成命題，並循環整合這些命題以形成命題網絡，由下往上建構一組原始、不連貫的心智模式。

2. 整合階段（integration stage）

整合階段，讀者將建構階段所形成的命題及命題網絡，結合相關的先備知識，開始不斷重複地進行整合，以形成涵蓋各鉅觀內容暨鉅觀命題間的因果與關係脈絡的知識結構，建立更具連貫、深層意義的情境模式。

3. 文本模式（textbase model）

文本模式均植基於文本成分，包括語法及語意，亦即微觀結構（microstructure）與鉅觀結構（macrostructure）。前者係指字、詞、句之間的關係，後者指涉文本大意暨段落間的關係。

4. 情境模式（situation model）

情境模式係為讀者將前階段所建構的文本模式，與先備知識進行整合，並且建構比表面字句理解更融貫且具深層意義的高階知識表徵。Rouet（2006）指出深層理解需要讀者整合文本訊息及本身先備知識的能力，亦即讀者會結合從文本新輸入的訊息與從記憶活化的先備知識，藉此形成心智表徵。

（二）多重文本閱讀理解歷程

多重文本的閱讀除了進行文本內的理解外，更連結跨文本間的訊息以發展對文本的情境認知（Wolfe & Goldman, 2005）。Bråten 等人（2009）指出闡釋多重文本理解的理論，以 Perfetti、Rouet 與 Britt（1999）所提出的文件模式影響現今最鉅，係為 CI Model 的延伸，包括文本間模式（intertext model）和情境模式，茲闡釋如下：

1. 文本間模式

文本間模式係為多重文本理解的重要元素，主要表徵文本間的意義，包括文本中的訊息，以及文本間訊息的關係與差異。Schwarz（2003）指出，多重文本的訊息整合能引導學生進行新知識的精緻化。

2. 情境模式

Bråten（2009）等人認為讀者整合自身長期記憶與各文本間的訊息後，推導出文本情境的解釋，即為情境模式。Rouet（2006）認為情境模式可因為新輸入的訊息而重新組織，故得以持續更新。Johnson 與 Seifert（1999）將情境模式的更新區分為表面更新與深層更新，前者係指讀者即使辨識出與先前知識抵觸的正確訊息時，也不會重新組織情境表徵；後者則會改變新輸入訊息所挑戰的表徵，致使情境模式更新。

二、多重文本閱讀理解評量架構

（一）多重文本閱讀理解的內涵

Bråten 與 Strømsø (2010) 指出多重文本素養係指個體針對特定議題、計畫或情境，建構一個整合並檢具意義的表徵時，所需定位、評估及應用不同訊息來源的能力。是以，多重文本理解涵蓋高階認知歷程與技能，還包括先備知識、來源、深層策略使用、任務覺察、文件專業知識，以及個人知識信念 (epistemology)。因此檢測可分成三個層次：其一，文本的表面理解；其二，文本的情境理解；其三，文本間訊息的推論能力，這部分主要檢測讀者擷取跨文本中的訊息，並加以整合以做出推論的能力。

（二）多重文本閱讀理解評量架構的界定

Wolfe 與 Goldman (2005) 指出讀者進行閱讀時，會建構心理表徵，包括表面字句的鋪陳、文章本身的意義暨詮釋模式 (情境模式)。另 van den Broek (1990) 提出因果推論模型，認為讀者會試圖尋求新訊息與舊訊息的連結，若有直接因果關係，則不需作推論，倘沒有因果關係，則會試圖連結情境模式既有的舊訊息或長期記憶儲存的資訊以產生推論，故推論為讀者形成融貫心理表徵的重要能力。Royer 等人 (1996) 針對文本閱讀理解評量，提出句子判斷技術 (sentence verification technique, SVT) 與推論判斷技術 (inference verification technique, IVT)。SVT 主要測量記憶文本內容要旨的能力，IVT 則是測量推論能力，其一為近推論，意指連結文本內二個訊息所作出的推論；其二為遠推論，係指連結文本訊息及先備知識所作出的推論。因此，文本表面意義的理解係為近推論與遠推論的必要條件。此外，Bråten 等人 (2010) 提出句子判斷作業 (sentence verification task)、文本內推論判斷作業 (intratextual inference verification task, IntraVT)，暨文本間推論判斷作業 (intertextual inference verification task, InterVT)。其中，句子判斷作業所評測的能力與 Royer 的 SVT 相同，都是測量記憶文本內容要旨的能力，IntraVT 則是測量單一文本的情境理解，而 InterVT 主要測量跨文本間的推論能力。基此，多重文本閱讀理解評量可從五個方面探討如下：

1. 先備知識

進行文本閱讀時，與文本相關的先備知識，可協助個體提取重要訊息 (Rouet, Britt, Mason, & Perfetti, 1996)。CI Model 的建構階段，文本訊息會激發長期記憶中相關的知識進行解讀，整合階段亦會植基先備知識建構融貫的情境模式。因此很多研究均會進行先備知識的前測，透過排除先備知識變項影響的統計控制，了解個體閱讀理解的程度。

2. 提取訊息 (information retrieval)

「提取訊息」係指透過長期記憶的激發與提取，進行文本字義與表面意義的理解，藉以了解文本所欲表達的想法和概念。換言之，讀者能從文本當中提取與閱讀目標相關的訊息，分辨文章整體訊息或主題，包括特定詞彙定義、主旨及傳達觀點之表面理解，屬於「直接理解」之淺層的知識處理歷程。

3. 概化訊息 (information generalization)

「概化訊息」係指將「提取訊息」階段所建立的微觀結構，抽取主要概念並整合成文本大意的鉅觀結構，所萃取的概念範疇均未超越文本內容。基此，讀者必須具備歸納訊息及重點的能力，裨益建立融貫的情境模式。

4. 解釋訊息 (information interpretation)

「解釋訊息」係指在沒有增加新訊息的前提下，重組文章主旨，有異於「整合訊息」會演繹內容訊息到理解的情境中。此階段必須活化先備知識，建構深層理解，推論文本所隱含的意義，並用自己的話表述文本訊息，以建構符合主旨的情境模式。

5. 整合訊息 (information integration)

學者 (Rouet, Vidal-Abarca, Erboul, & Millogo, 2001) 指出情境模式的建構包括解析文字內涵、建構微觀及鉅觀命題、產生推論，暨透過先備知識整合文本訊息。讀者在「整合訊息」階段會演繹內容訊息到理解情境，表達對文本正向或負向的判斷 (Wolfe & Goldman, 2005)，並對照比較文本間的訊息，進行文本內容的評鑑理解與批判理解，亦即能演繹觀點，進行文本內容的分析，係為涵蓋鉅觀命題內容、因果及關係脈絡之情境模式。

綜上所述，本研究認為多重文本閱讀理解的歷程，首先會進行字詞解讀與句型解析，以形成對文本表面意義的理解，此即 Kintsch (1988) 所謂微觀結構，屬於「提取訊息」；接著抽取主要概念，加以整合成命題網絡並建立一個表面的意義描述，亦即文本大意及段落間關係的鉅觀結構，其概念範疇均未超越文本所提及的概念，此即為「概化訊息」；再者，將前二階段所建構的命題網絡，與先備知識進行整合，在沒有演繹新訊息的前提下，推論文本所隱含的意義，建構對文本深層理解的情境模式，此即為「解釋訊息」；最後則演繹觀點至理解的情境中，表達對文本正向或負向的判斷，藉以對文本內容進行評鑑理解與批判理解，係為「整合訊息」。

三、多面向 Rasch 測量模式

多面向 Rasch 測量模式 (Many-Facet Rasch Measurement, 以下簡稱 MFRM) 是 Rasch 測量模式的延伸 (Eckes, 2009)，適用於多個潛在變項影響評量結果的測量模型，暨評分者中介的評量。因此本節將依序介紹 Rasch 測量模式及延伸模式，包括多元計分模式、評定量尺模式 (rating scale model, 以下簡稱 RSM)、部分給分模式 (partial credit model, 以下簡稱 PCM)，以及 MFRM 測量模式。

(一) Rasch 測量模式

Rasch 模式認為作答反應會受到試題難度與受試者能力的影響，二元計分 (dichotomous scoring) 的數學模式，係指作答反應非對即錯的差別，亦即 0 分與 1 分的區別，其數學模式如 1.1 公式。

$$\ln \left[\frac{p_{ni1}}{p_{ni0}} \right] = \theta_n - \beta_i \quad (1.1)$$

θ_n ：受試者 n 的能力估計值； β_i ：第 i 題的難度估計值； p_{ni1} ：受試者 n 在第 i 題答對的機率； p_{ni0} ：受試者 n 在第 i 題答錯的機率。

(二) Rasch 延伸模式

1. 多元計分模式

多元計分 (polytomous scoring) 係指作答反應有多個計分規則，亦即並非只有 0 分與 1 分的區別，而是存在很多計分，其數學模式如 1.2 公式。

$$\ln \left[\frac{p_{nij}}{p_{ni(j-1)}} \right] = \theta_n - \beta_{ij} \quad (1.2)$$

θ_n ：受試者 n 的能力估計值； β_{ij} ：第 i 題在第 j 階的難度估計值； p_{nij} ：第受試者 n 在第 i 題得 j 分的機率； $p_{ni(j-1)}$ ：受試者 n 在第 i 題得 $(j-1)$ 分的機率。

2. 評定量尺模式

有 k 個點數 (1, 2, ..., k-1) 的多分題，就具有 $k-1$ 個閾難度 (threshold difficulty)，亦即從 1 分到 2 分、...k-1 分到 k 分的閾難度。RSM 即假設所有題目的閾難度都相同，換言之評量所有題目都使用相同的評定量尺，即適用 1.3 公式。

$$\ln \left[\frac{p_{nij}}{p_{ni(j-1)}} \right] = \theta_n - (\beta_i + C_j) \quad (1.3)$$

θ_n ：受試者 n 的能力估計值； β_i ：第 i 題的難度估計值； C_j ：第 j 階閾的難度估計值（第 $j-1$ 階進到第 j 階所增加的難度）； p_{nij} ：受試者 n 在第 i 題得 j 分的機率； $p_{ni(j-1)}$ ：受試者 n 在第 i 題得 $(j-1)$ 分的機率。

3. 部分給分模式

PCM 即假設多分題模式的閾難度不同，亦即評量所有題目使用不同的評定量尺，即適用 1.4 公式。

$$\ln \left[\frac{p_{nij}}{p_{ni(j-1)}} \right] = \theta_n - (\beta_i + C_{ij}) \quad (1.4)$$

θ_n ：受試者 n 的能力估計值； β_i ：第 i 題的難度估計值； C_{ij} ：第 i 題，第 j 階閾的難度估計值（第 $j-1$ 階進到第 j 階所增加的難度）； p_{nij} ：受試者 n 在第 i 題得 j 分的機率； $p_{ni(j-1)}$ ：受試者 n 在第 i 題得 $(j-1)$ 分機率。

4. MFRM 測量模式

(1) MFRM 的數學公式

Linacre (1989) 指出 MFRM 係為 Rasch 模型所延伸發展而來的測驗模式，包含多個面向 (facets)，如受試者能力、試題難度及評分者嚴苛度等。此模型系將所有需考量的潛在面向放在同一個尺度上，並將分數轉化為一等距尺度 logit 分數（參照公式 1.5）。

$$\ln \left[\frac{p_{nij}}{p_{ni(j-1)k}} \right] = \theta_n - (\beta_i + C_{ij} + \eta_k) \quad (1.5)$$

θ_n ：受試者 n 的能力估計值； β_i ：第 i 題的難度估計值； C_{ij} ：第 i 題，第 j 階閾的難度估計值（第 $j-1$ 階進到第 j 階所增加的難度）； η_k ：評分者 k 的嚴苛度； p_{nij} ：受試者 n 被評分者 k 在第 i 題評 j 分的機率； $p_{ni(j-1)k}$ ：受試者 n 在第 i 題被評分者 k 在第 i 題評 $(j-1)$ 分的機率。

(2) MFRM 的分析應用

有鑑於很多測驗情境，並非只有受試者能力及試題難度會影響其作答反應，以建構反應試題為例，即便提供客觀一致的評量規準，評分者仍會有主觀的判斷，因此產生除能力及難度因素外，也會影響作答反應的其它因素 (Landy & Farr, 1983)。Linacre (1989) 研究指出，影響作答反應主要有四個因素，除了受試者能力及試題難度外，還包括評分者嚴苛度及評分者心中的那把評定量尺。基此，MFRM 就被用來解決這類的問題。證據指出 (Eckes, 2009)，評分者判斷存在相當程度之系統性錯誤，若不加解釋，極可能導致不適當或不公平的結論。是以，如果忽略相關影響受試者反應表現的潛在面向 (facets)，可能會誤導或誤判評測結果，致使造成能力估計的偏誤。因此雙重 Rasch 測量模式延伸的 MFRM，即可考量能力及難度以外更多的面向，如嚴苛度、任務難度、計分規準等潛在面向以進行分析，藉以客觀估計能力、嚴苛度、任務及規準難度，以及面向間潛在交互作用的偵測。另進行 MFRM 分析的資料不限定二元計分，可以是多元計分資料，尤其是評分者中介的評量。

方法

一、研究架構

本研究首先召集 2 位研究專家教師及 2 位大學專家教師，討論並發展「科學多重文本閱讀理解評量」，以及「多重文本閱讀理解評量規準」。緊接著預試，並將資料進行項目分析與探索性因素分析，藉以修訂「科學多重文本閱讀理解評量」及「多重文本閱讀理解評量規準」。修訂完成即正式施測，並回收資料與進行閱卷。最後進入分析階段，包括建構效度與信度的檢驗，前者係進行二階驗證性因素分析，透過考驗理論模式與實徵資料的適配程度，藉此確立評量構念是否能為實徵資料所支持而檢具建構效度；後者則依序檢驗評分者一致性與評量試題內部一致性。評分者部分，研究者分別進行古典測驗理論之評分者一致性分析，及現代測驗理論 MFRM 之評分者嚴苛度的估計、RSM 及 PCM 模式的卡方考驗與貝氏訊息準則（Bayesian Information Criterion, BIC）的模式比較，藉以檢核評分者的一致性（consistency）和變異性（variability）。另進行評量試題之內部一致性的估計，以檢視評量的信度。

二、研究對象

（一）「科學多重文本閱讀理解評量」預試暨正式施測對象

本研究以 5~9 年級共計 165 人進行預試，透過項目分析與探索性因素分析，修訂評量與規準，並提供各向度之內容水準所對應表現水準的範例。另以 5~9 年級共計 1535 人進行正式施測。為平衡文本順序的影響，將 2 篇文本順序分為 A、B 二版，受試者隨機接受 A 或 B 版本，亦即有一半讀者接受 A 版本（氣候變遷與三峽大壩有關 Vs. 氣候變遷與三峽大壩無關），另一半讀者則接受 B 版本（氣候變遷與三峽大壩無關 Vs. 氣候變遷與三峽大壩有關）。

（二）「科學多重文本閱讀理解評量」之閱卷老師

本研究的評分者包括研究者共計有 3 位，另外二位分別是資訊工程及健康促進背景的博士級研究員。評閱工作開始前，評分者需針對評分規準進行討論及訓練，之後才開始進行評閱工作。為避免評分者因工作負荷大導致倦怠而影響評分品質，本研究透過等化設計，將 3 位評分者的評分結果轉換到同一量尺上，以利比較。基此，本研究設計評分者間 31%至 38%的定錨試卷，以便作為評分結果間的連結之用。

三、科學多重文本閱讀理解評量預試題本（SMTRCA）之編制

（一）評量範疇

本評量命題係以「氣候變遷與中國長江三峽大壩的關係」之科學議題探討為主，涵蓋氣象、物理、地質、地球科學及生態環境等領域。

（二）題本發展

本評量之閱讀文本以題組為單位，每一題組均由一篇文本及一組試題組成。SMTRCA 由 2 篇文本組成，文體為說明文，係以 Hartman 與 Allison（1996）所提出補充性文本及衝突性文本的組織方式，針對「氣候變遷與三峽大壩之間的關係」分別提出三峽大壩是否造成氣候變遷二個對立見解觀點的文本，內容介於 983~1000 個字，（參照表 1、表 2）。

（三）試題發展

SMTRCA 由 2 組試題組成，涵蓋「理解文本的能力」及「詮釋文本的能力」。前者包括「提取訊息」及「概化訊息」，以選擇題呈現，每一個文本均發展 6 題選擇題，共計 12 題反應試題，藉

以檢視學生「理解文本的能力」。後者則包括「解釋訊息」及「整合訊息」，以建構題呈現，共發展 7 題文本內暨 2 題跨文本之建構題，合計 9 題建構試題。每一題建構題均提供清楚、聚焦陳述的回答提示，藉以激發學生的先備知識與既有基模，致使其能進行修辭、概念及語意觀點之解釋與推論的動態心智活動，並進一步組織回答任務。藉以檢視學生「詮釋文本的能力」。試題架構茲詳述如下：

表 1 「氣候變遷與三峽大壩之間的關係」文本內容架構

文本訊息		內容	數量
關鍵名詞	科學名詞	氣象 氣流，溫度，濕度，風霧，降雨，乾旱，旱災，天氣，低氣壓，冷空氣，全球暖化，氣候變遷，氣候異常，氣候反常，氣候調節，循環交替，極端氣候，氣溫失衡，降雨分布，大氣循環，大氣環流異常，反聖嬰現象	23
		物理 蒸發，凝結，降水，水汽，水氣，水循環，物理變化	7
		地質 河床，河道，水位，人工屏障	4
		地球科學 大氣，切蝕，水流量，暴漲期，枯水期，暖濕氣流，木桶效應，大氣循環，冷暖交會，攔腰截斷，截流積聚，太平洋水氣	12
		生態 生態環境，水量平衡	2
	一般名詞	重慶，四川，湖北，水庫，蓄水，江面，洞庭湖，鄱陽湖，印度洋，太平洋，三峽大壩，湖泊水面，長江流域，四川盆地，長江中下游，青康藏高原	16
關鍵概念	氣象	寬闊水面使水分蒸發範圍變大，形成低氣壓，影響大氣循環。	9
		流動江水會將熱量帶走，對周圍環境產生天然氣候調節的作用。	
		降雨天氣是由暖濕氣流與冷空氣在高空發生激烈碰撞而形成。	
		極端氣候是指氣候狀態嚴重偏離平均態，屬於不易發生的事件。	
		大氣中暖濕氣流無法與冷空氣形成冷暖交會，因而不易降雨。	
		暖濕氣流始終不強盛，不利形成降雨，因此造成乾旱現象。	
		反聖嬰現象導致太平洋水氣難以向長江中下游地區輸送。	
		水汽不足，缺乏降雨繼而導致乾旱氣候的發生。	
	物理	暖濕氣流不強盛，不利形成降雨。	4
		水循環指水以蒸發、凝結、降水等物理變化，進行周而復始的循環。	
		江面和流域內的湖泊水面面積急遽減少，使得水氣蒸發量減少。	
		水分蒸發量範圍變大，局部地區降水量增加。	
	地質	靜態儲水，喪失攜帶熱量的功能。	4
		三峽大壩阻斷天然河道，導致大量泥沙淤積在水庫，因而使水庫下泄不帶泥沙的清水。	
		水庫攔阻泥沙，下泄清水切蝕河床，造成河道水位下降。	
		200 公尺高崖的三峽大壩，儼然成一道人工屏障，阻擋長江江面的水氣正常進出四川盆地。	
論述觀點	氣候變遷	河道水位下降，使得長江中下游洞庭湖及鄱陽湖的湖水流入長江，導致兩湖水位大減而乾涸。	3
		江水被截流積聚在大壩庫區，管控限制下泄水流量，影響水量平衡。	
	地域特徵	海洋與陸地之間的水循環稱為外循環。	2
		局部區域內的水循環稱為內循環。	
論述觀點	氣候變遷	旱災成因是因為大氣環流異常及全球暖化，所導致極端氣候的現象。	2
		三峽大壩工程，實則為氣候變遷的殺手。	
論述觀點	地域特徵	長江中下游地區由於暖濕氣流始終不強盛，不利形成降雨，因此造成乾旱現象。	2
		儼然成為人工屏障，阻擋水氣進出四川盆地，導致乾旱氣候的發生。	

表 1 (續)

論 述 觀 點	氣候調節	三峽大壩局部地區的水循環，相較於印度洋和太平洋大範圍區域的水循環，對氣候影響極其有限。 從青康藏高原下來的流動雪水被三峽大壩攔腰截斷，成為靜態儲水，喪失攜帶熱量的功能。
	水量控管	三峽大壩每年在長江暴漲期囤蓄水量，並在枯水期釋放存水，為長江中下游補水解旱。 江水被三峽大壩攔腰截斷，從而管控限制下泄的水流量，造成長江中下游乾旱。
	降雨分布	反聖嬰現象，導致暖濕氣流無法在長江流域與冷空氣冷暖交會，因而不易降雨，另全球暖化造成雨量驟減及降雨量不均而導致乾旱。 三峽大壩導致長江中下游的江面和湖泊水面面積減少，使得水氣蒸發量減少，改變降雨分布，引發乾旱。
	水文地質	三峽大壩下泄的水流量，至少比天然水量高出，能有效支持長江中下游抗旱。 三峽大壩阻斷泥沙淤積而下泄清水，長年切蝕中下游的河床，造成河道水位下降，使得洞庭湖及鄱陽湖湖水流入長江，導致湖泊乾涸。

表 2 「氣候變遷與三峽大壩之間的關係」文本爭議觀點比較

文本訊息			氣候變遷與長江三峽大壩有關	氣候變遷與長江三峽大壩無關
關鍵名詞	科學名詞	氣象	21	29
		物理	6	14
		地質	9	0
		地科	8	10
		生態	4	1
	一般名詞		35	33
合計		83	87	
關鍵概念	科學概念	氣象	3	6
		物理	3	1
		地質	4	1
		水文	1	2
	合計		11	10
見解觀點	氣候變遷		1	1
	地域特徵		1	1
	氣候調節		1	1
	水量控制		1	1
	降雨分布		1	1
	水文地質		1	1
合計		6	6	

1. 告知學生應該根據文本內容，就其所提出的問題進行回答。

例題：請問長江中下游的洞庭湖及鄱陽湖等湖，為何會萎縮乾涸？

2. 針對文中作者所表述的具體說明及重要訊息，請學生回答作者關於主題探討的觀點及理由，並表述個人對於文中作者觀點的贊同與否定的理由。

例題 1：請問本文作者是否認為：長江三峽大壩工程是導致中國西南及長江中下游地區連年發生旱災的原因？並請敘述作者所持理由。

例題 2：你認為本文作者所敘述長江地區連年發生旱災與長江三峽大壩工程關係的觀點，合理？還是不合理？

3. 針對由二篇科學概念論述相互關聯，但議題探討觀點則持對立見解的文本，請學生綜合以上二篇科學文本，提出個人贊同或否定論點，同時就個人所支持的見解，表述理由。

例題：讀完以上二篇科學文章，你認為長江三峽大壩工程是否是導致中國西南及長江中下游地區連年發生旱災的原因？請提出你個人贊同或不贊同的理由。

（四）評量格式與評量解釋

1. 評量格式

本評量包括選擇題與建構題，評量結果採用標準參照暨有診斷功能之分析式模式，提供提取訊息、概化訊息、解釋訊息，及整合訊息等向度的表現水準，其中整合訊息又分為文本內及跨文本之整合訊息的能力。

2. 評量解釋

本評量之選擇題採用二元計分，答對得 1 分，答錯得 0 分；建構題則採用多元計分，依據「多重文本閱讀理解評分規準」，評閱作答表現，包括「進階」、「精熟」、「基礎」、「待加強」，依序評定為 4、3、2 及 1 分（參照表 3），未作答，則記為“.”。其中，文本內訊息之整合能力包括 C4~C9 共計 4 題，跨文本訊息之整合能力包括 C8~C9 共計 2 題（參照表 6）。

表 3 「科學多重文本閱讀理解評量」評分標準

向度	試題型態	計分型態	受試反應	分數	評分指標
提取訊息	選擇題	二元計分	答錯	0	解讀文本關鍵字詞 解析文本關鍵句型
			答對	1	提取閱讀文本目標相關的知識與記憶 理解文本關鍵語意以及主旨大綱概念
概化訊息	選擇題	二元計分	答錯	0	抽取文本主要概念 歸納文本主要訊息
			答對	1	摘要文本主要重點 建立文本大意及段落關係的鉅觀結構
解釋訊息	建構題	多元計分	無法正確地	1	萃取命題網絡並與先備知識進行整合
			能夠粗略地	2	建構符合全文主旨的命題之情境模式
			能夠正確地	3	能用自已的話表述文本所傳達的訊息
			能正確及完整地	4	推論文本隱含意義
整合訊息	建構題	多元計分	無法正確地	1	對照比較文本訊息並與先備知識整合
			能夠粗略地	2	建構涵蓋鉅觀命題之融貫的情境模式
			能夠正確地	3	依據文本傳達訊息演繹出自己的觀點
			能正確及完整地	4	對文本內及文本間據以進行判斷分析

四、多重文本閱讀理解評量規準（RMTRCA）之發展

多重文本閱讀理解評分規準係指明確闡釋內容標準、表現標準，以及所欲評估能力表現之特點或範疇的指引，用以闡釋學生「應知」與「應能」的水準程度，茲分別簡述評分規準制訂歷程。

（一）釐清評量目的

為避免主觀評分所導致回饋的變異，致使評分的客觀性受到質疑，本研究發展 RMTRCA，以克服人為評量的主觀偏誤。

（二）訂定表現標準個數暨命名類別

本研究依據美國國家教育進展評量（National Assessment of Educational Progress, NAEP）所提出之表現標準標籤命名藍圖，將 RMTRCA 之表現標籤水準（Performance Label Levels, PLLs）命名為基礎、精熟及進階，進而將學生區分成待加強、基礎、精熟及進階四個表現類別。

（三）建置表現標準的陳述

RMTRCA 之內容標準包括「提取訊息」、「概化訊息」、「解釋訊息」及「整合訊息」四個向度。另表現水準依序為待加強、基礎、精熟及進階，並擬定內容標準暨表現水準的陳述（參照表 4）。

五、評量內容適切性之評估

本研究選擇 5~9 年級學生進行預試，以檢視受試者是否理解評量試題之語意結構，並評估作答時間。同時邀請閱讀理解領域大學教授、可讀性系統研發之資訊工程及國家災害防救科技中心研究員，計 4 位專家，依據預試結果，提供修正與補充意見，最後確定 SMTRCA 共計 21 題。

六、預試分析與結果

（一）項目分析及結果

本研究將預試資料進行項目分析，並依據以下刪題標準進行題目的篩選或修訂：（1）同質性考驗：計算各題和全量表的相關， $< .3$ 者刪除；（2）一致性考驗：刪除該題可提高 *Cronbach's α* 值者，予以刪除；（3）探索式因素分析（exploratory factor analysis, EFA）之主軸因素法（principal axis factoring, PAF）因素負荷量： $< .3$ 者刪除；（4）估計決斷值（CR: critical ration）：未達顯著者刪除。分析結果顯示，「提取訊息」之第 1、2 題未符合上述篩選標準，故予以刪除，共刪除計 2 題。

表 4 多重文本閱讀理解評量規準

表現類別		表現標準敘述		
水準	計分	提取訊息	概括訊息	解釋訊息
待加強	1	●無法正確地提取與文本訊息相關的知識與記憶。 ●無法正確地解讀文本字詞。 ●無法正確地解析文本句型。 ●無法正確地理解文本語意及主旨概念。	●無法正確地抽取文本主要概念。 ●無法正確地歸納文本主要訊息。 ●無法正確地摘要文本主要重點。 ●無法正確地建構文本大意及段落間關係的鉅觀結構。	●無法正確地對照及比較文本間的訊息，與先備知識進行整合。 ●無法正確地建構融貫的情境模式。 ●無法適當地依據文本所傳達的訊息，演繹自己的觀點。 ●無法適當地針對文本內及文本間的訊息，進行判斷與分析。
		●能部分地提取與文本訊息相關的知識與記憶。 ●能部分地解讀文本字詞。 ●能部分地解析文本句型。 ●能部分地理解文本字句語意及主旨概念。	●能部分地抽取文本主要概念。 ●能部分地歸納文本主要訊息。 ●能部分地摘要文本主要重點。 ●能部分地建構文本大意及段落間關係的鉅觀結構。	●能粗略地對照及比較文本間的訊息，與先備知識進行整合。 ●能粗略地建構融貫的情境模式。 ●能粗略地依據文本所傳達的訊息，演繹自己的觀點。 ●能粗略地針對文本內及文本間的訊息，進行判斷與分析。
精熟的	3	●能正確地提取與文本訊息相關的知識與記憶。 ●能正確地解讀文本字詞。 ●能正確地解析文本句型。 ●能正確地理解文本語意及主旨概念。	●能正確地抽取文本主要概念。 ●能正確地歸納文本主要訊息。 ●能正確地摘要文本主要重點。 ●能正確地建構文本大意及段落間關係的鉅觀結構。	●能正確地對照及比較文本間的訊息，與先備知識進行整合。 ●能正確地建構融貫的情境模式。 ●能適當地依據文本所傳達的訊息，演繹自己的觀點。 ●能適當地針對文本內及文本間的訊息，進行判斷與分析。
		●能正確地及完整地提取與文本訊息相關的知識與記憶。 ●能正確地及完整地解讀文本字詞。 ●能正確地及完整地解析文本句型。 ●能正確地及完整地理解文本語意及主旨概念。	●能正確地及完整地抽取文本主要概念。 ●能正確地及完整地歸納文本主要訊息。 ●能正確地及完整地摘要文本主要重點。 ●能正確地及完整地建構文本大意及段落間關係的鉅觀結構。	●能正確地及完整地對照及比較文本間的訊息，與先備知識進行整合。 ●能正確地及完整地建構融貫的情境模式。 ●能適當地及完整地依據文本所傳達的訊息，演繹自己的觀點。 ●能適當地及完整地針對文本內及文本間的訊息，進行判斷與分析。

(二) 探索性因素分析及結果

本評量共計 19 個觀察變數，依序為「提取訊息」(S7, S8, S9)、「概化訊息」(S3, S4, S5, S6, S10, S11, S12)、「解釋訊息」(C1, C2, C3)、「整合訊息」(C4, C5, C6, C7, C8, C9)。

1. KMO 和 Bartlett's 球形檢定

本研究以 KMO 和 Bartlett's 球形檢定作為判定資料是否適合進行因素分析的依據，結果顯示 KMO 為 .92，Bartlett's 球形檢定達顯著差異，表示資料適合進行因素分析。

2. 因素相關係數估計

相關矩陣結果顯示，本評量題項間的相關介於 .35~.70，呈現相當程度的相關。

3. 因素負荷量估計

本研究萃取四個因素，以主軸因素法暨最小斜交法 (direct oblimin) 進行因素分析，所抽取 4 個因素間的相關係數介於 .33~.90，解釋變異量為 46.33% (參照表 5)。

4. 因素命名

(1) 因素 1

因素 1 特徵值為 34.95，組型負荷量絕對值依序 C4 為 .60、C5 為 .68、C6 為 .88、C7 為 .90、C8 為 .66、C9 為 .72，共 6 題，命名「整合訊息」。

(2) 因素 2

因素 2 特徵值為 6.03，組型負荷量絕對值依序 S3 為 .33、S4 為 .54、S5 為 .39、S6 為 .44、S10 為 .39、S11 為 .49、S12 為 .60，共 6 題，命名為「概化訊息」。

(3) 因素 3

因素 3 特徵值為 2.92，組型負荷量絕對值依序 S7 為 .53、S8 為 .57、S9 為 .57，共 3 題，命名為「提取訊息」。

(4) 因素 4

因素 4 特徵值為 2.43，組型負荷量絕對值依序 C1 為 .34、C2 為 .82、C3 為 .75，共 3 題，命名為「解釋訊息」。綜上所述，依據項目分析與因素分析結果顯示評量各題項檢測之因素符合理論構念，組型負荷量絕對值皆>.30，提供評量構念效度之參酌證據。

表 5 SMTRCA 轉軸後之組型矩陣

題項構念	代號	斜交轉軸後之組型負荷量				
		因素1	因素2	因素3	因素4	共同性 h^2
整合訊息	C7	.90	.01	-.03	.09	.69
	C6	.88	-.03	.04	.04	.73
	C9	.72	-.01	-.04	-.13	.63
	C5	.68	.05	.05	-.08	.62
	C8	.66	-.01	.01	-.17	.62
	C4	.60	.12	.15	-.06	.62
概化訊息	S12	.07	.60	-.04	.05	.35
	S4	.04	.54	.04	-.05	.36
	S11	.03	.49	-.10	-.00	.21
	S6	-.03	.44	.13	-.07	.30
	S10	.02	.39	.15	.07	.22
	S5	-.06	.39	.06	-.07	.18
	S3	.01	.33	.00	-.06	.14
	S8	.04	-.04	.57	-.02	.33
提取訊息	S9	.05	.17	.57	.05	.45
	S7	.02	-.01	.53	-.06	.31
	C2	.05	.05	.07	-.82	.82
解釋訊息	C3	.15	.07	-.01	-.75	.79
	C1	.24	.15	.06	-.34	.44
總累計變異量		46.33%				

七、正式評量題本之分析結果

(一) 正式評量題本之構念與題目

本評量經項目分析與因素分析後刪除 2 題，保留 19 題，依序為「提取訊息」3 題、「概化訊息」7 題、「解釋訊息」3 題，以及「整合訊息」6 題，其中「整合訊息」包括「文本內的整合 (intratextual integration)」4 題及「跨文本的整合 (intertextual integration)」2 題（參照表 6）。

(二) 計分標準與實施細則

1. 計分標準

本評量之「提取訊息」與「概化訊息」採用二元計分模式，答對得 1 分，答錯則得 0 分，前者得分範圍介於 0~3 分，後者則介於 0~7 分。另「解釋訊息」與「整合訊息」則採多元計分模式，評分者根據讀者答題表現進行評分，總分為 4 分，評分層次包括：(1) 無法正確地建構，計 1 分；(2) 能粗略地建構，計 2 分；(3) 能正確地建構，計 3 分；(4) 能正確地及完整地建構，計 4 分，前者得分範圍介於 1~12 分，後者則介於 1~24 分。

2. 施行細則

本評量採團體施測，每位受試者均有一份評量題本及答案紙，作答時間計 85 分鐘，加上指導說明，共需 90 分鐘完成。首先請受試者填寫基本資料，其次進行作答說明，最後確認受試者了解作答方式後，才正式進行施測。

表 6 SMTRCA 之向度與檢測能力摘要表

評量向度	主要檢測能力	題號	題數
提取訊息	提取閱讀文本目標相關的知識與記憶； 解讀文本關鍵字詞； 解析文本關鍵句型； 理解文本關鍵語意及主旨大綱概念。	S7, S8, S9	3
概化訊息	抽取文本主要概念； 歸納文本主要訊息； 摘要文本主要重點； 建立文本大意及段落關係的鉅觀結構。	S3, S4, S5, S6, S10, S11, S12	7
解釋訊息	淬取命題網絡並與先備知識進行整合； 建構符合全文主旨的命題之情境模式 推論文本隱含意義； 能用自己的話表述文本所傳達的訊息。	C1, C2, C3	3
整合訊息	對照比較文本訊息並與先備知識整合； 建構涵蓋鉅觀命題之融貫的情境模式； 依據文本傳達訊息演繹出自己的觀點； 對文本內及文本間據以進行判斷分析。	文本內 C4, C5 C6, C7	4
		跨文本 C8, C9	2
合 計			19

（三）資料處理與統計分析

本研究施測資料回收後，首先請評分者依據評量規準進行評分，最後將資料鍵入電腦，依序進行項目分析、評分者一致性與評量內部一致性的檢驗，以及二階驗證性因素分析。

結果

一、信度檢驗

（一）評分者內的一致性

本研究將評分者評閱結果，依據解釋訊息、整合訊息、文本內整合訊息、跨文本整合訊息，以及詮釋文本等向度的試題分布，分別加總起來，進行評分者內一致性的分析。結果顯示 3 位評分者內之 *Cronbach's α* 值均為 .91，顯示評分者內的一致性相當穩定。

（二）評分者間的一致性

1. Kendall ω 和諧係數分析

如表 7 所示，3 位評分者共同評閱 32 位學生。分析結果顯示「解釋訊息」、「整合訊息」、「文本內整合訊息」及「跨文本整合訊息」4 個向度分評量的 *Kendall ω* 和諧係數值均為 .80 以上，卡方值（ χ^2 ）依序為 85.20、82.45、76.89、79.27，卡方考驗達顯著（ $p < .001$ ），統計考驗拒絕虛無假設，並接受對立假設，顯示評分者間有顯著相關存在，亦即評分者間相對等級的評分趨勢相當一致。

表 7 Kendall ω 和諧係數分析結果 ($N = 32$, $df = 31$)

	解釋訊息	整合訊息	文本內整合	跨文本整合
Kendall ω	.920	.890	.830	.850
χ^2	85.200	82.450	76.890	79.270
P	.000	.000	.000	.000

2. 評分者嚴苛度一致性的考驗

(1) 卡方考驗

a. MFRM 之卡方考驗

本研究進行 MFRM 二個 facet 分析 (rater+criteria)，結果顯示分散指標信度 (separation reliability) 為 .97，卡方考驗達顯著 ($p < .001$)，表示評分者間的嚴苛度確實存在差異 (參照表 8、圖 2)，其嚴苛度差異廣度介於 .082~.506 logits。符合 Rasch 模式中 facets 均相互獨立的基本假定，亦即每個評分者應具有獨特的嚴苛度。此外，infit 與 outfit MFRM 均介於 1 ± 0.3 之間 (Eckes, 2005)，表示無論是嚴格或寬鬆的評分者，均能有效區分出高、低能力的學生。

表 8 MFRM 之評分者嚴苛度估計

facet	評分者	嚴苛度	ERROR	INFIT MFRM		OUTFIT MFRM	
				MFRM	T	MFRM	T
rater+criteria	A	-.310	.059	1.17	2.9	1.19	3.4
	B	.196	.057	1.12	2.2	1.12	2.2
	C	.114*		1.16	2.8	1.12	2.4
Separation Reliability = .97							
Chi-square test of parameter equality = 39.64, $df = 2$, Sig Level = 0.000							

b. RSM 及 PCM 之卡方考驗

有鑑於能力、嚴苛度，以及試題、任務或計分標準等難度均為 facets，可在同一個量尺上進行比較，故本研究將評分者嚴苛度視為試題難度，透過 RSM 及 PCM 模式比較，進行卡方考驗，藉以檢驗評分者間的閾值嚴苛度是否相同。結果顯示，RSM 與 PCM 的 χ^2 差值為 219.81，自由度差值為 28，卡方考驗達顯著水準，顯示實徵資料比較適配 PCM，亦即評分者間的閾值嚴苛度存在差異 (參照表 9)。

c. BIC 準則

本研究將 RSM 與 PCM 所估計的 Deviance 進行 BIC 的轉換，其數學模式如 1.6 公式。結果顯示兩個模式之 BIC 值差不多，但以 PCM 更為適配，表示評分者間存在不同評分的閾值嚴苛度 (參照表 9)。

$$-2\ln L + K \cdot \ln M \quad (1.6)$$

L：概似函數 (likelihood function)；K：模型參數個數；M：樣本數。

表 9 RSM 及 PCM 卡方考驗一覽表 ($N = 1535$)

		Deviance	轉換 BIC
Deviance (χ^2)	rating scale model	17566.36	17629.18
	partial credit model	17346.56	17507.10
df	rating scale model	18.00	18.00
	partial credit model	46.00	46.00
	χ^2 差值 ($\Delta\chi^2$)	219.81	122.08
	自由度差值 (Δdf)	28.00	28.00
卡方考驗結果		< .01, 達顯著水準	< .01, 達顯著水準

(三) 內部一致性信度

內部一致性分析結果顯示，全評量之 *Cronbach's α* 為 .90，各分評量之 α 值介於 .66~.91，依序「提取訊息」為 .66、「概化訊息」為 .68、「解釋訊息」為 .83、「整合訊息」為 .91。其中「提取訊息」、「概化訊息」之 *Cronbach's α* 值 < .7，但由於分量表題數較少，故每個構念之 α 值會較總評量低。基此，本評量之 *Cronbach's α* 信度尚在可接受的範圍。

二、效度檢驗

(一) 驗證性因素分析

本研究以最大概似法 (*ML*) 進行二階驗證性因素分析，藉以考驗 SMTRCA 之理論構念模式是否與實徵資料適配。建構信度分析結果顯示，一階因素之建構信度 ($\rho_c = (\sum \lambda)^2 / [(\sum \lambda)^2 + (\sum \varepsilon)]$) 依序：「提取訊息」為 .83、「概化訊息」為 .77、「解釋訊息」為 .91、「整合訊息」為 .94，二階因素之組合信度 (composite reliability, *CR*) 為 .91，表示建構信度尚佳。

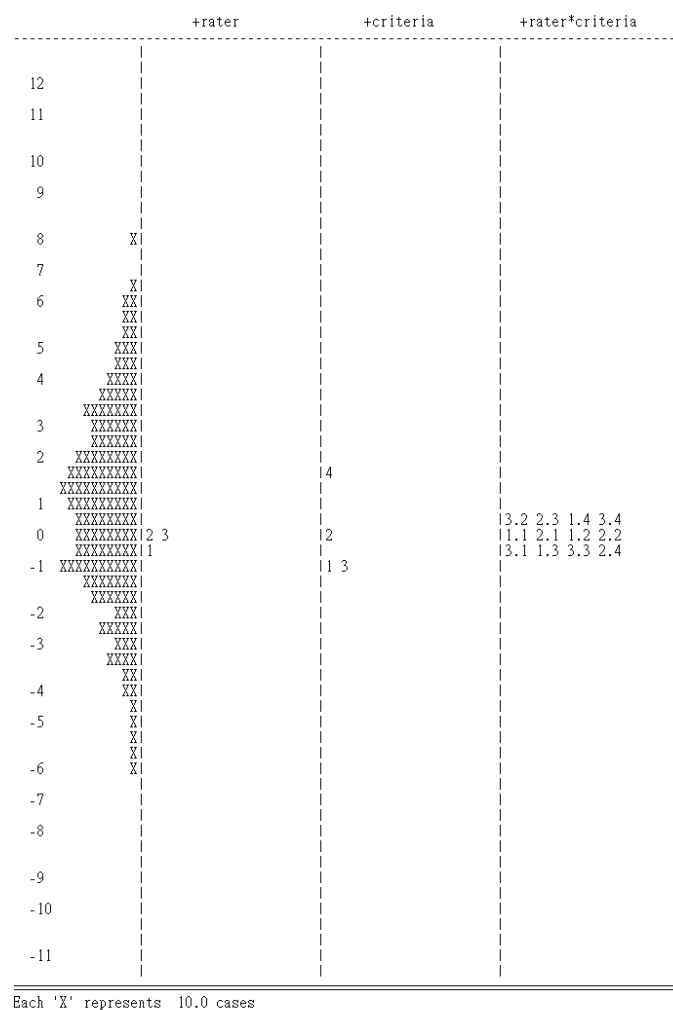


圖 2 MFRM 之變項分布圖

1. 模式辨識

本評量共計 19 個觀察變項，可提供 190 個資料點，實際估計參數為 42，自由度 148，整體模式可辨識。

2. 模式基本適配度的檢驗

模式基本適配度的檢驗結果顯示，各觀察變項之誤差皆無負值且達顯著水準；另因素負荷量介於 .49~.90 之間；再者迴歸加權係數均 >1.96 ，表示模式徑路係數均達 .05 之顯著水準，因此檢具存在價值（參照表 10）。

3. 模式整體適配度的檢驗

模式整體適配度的檢驗，結果顯示卡方分配達顯著水準（ $\chi^2 = 2950.15$ ， $p < .001$ ）表示假設模式與觀察值存在顯著差異。絕對適配指標，除 *GFI*、*RMSEA* 外，*RMR* 達理想適配之臨界值；另增值適配指標之適配指數則未盡理想；其次，精簡適配指標除卡方自由度比（ χ^2/df ）未盡理想外，*PNFI*、*PCFI* 皆在理想範圍。綜上所述，SMTRCA 理論構念模式與實際觀察值尚且符合，但未來

仍應持續蒐集資料，進行模式的驗證與修正，俾使理論構念模式檢驗之適配指標能夠達到理想適配水準（參照表 10）。

表 10 驗證性因素分析適配度評鑑（ $N = 1535$ ）

	評估項目	適配標準或臨界值	徑路模式檢定結果	模式適配評估
基本適配	誤差變異	1. 無負值	是	佳
		2. 達顯著水準	是	佳
	因素負荷量	介於 .5~.95	.49~.94 之間	佳
	CR 值	> 1.96，達顯著水準	是	佳
模式外在品質	χ^2 值	不顯著 ($p > .05$)	$\chi^2 = 2950.15$ ($p = .00$)	未盡理想
	絕對適配指標	$GFI \geq .90$	.83	尚可
		$RMR \leq .05$	.05	佳
		$RMSEA \leq .08$	.11	未盡理想
		$AGFI \geq .90$	.78	未盡理想
	增值適配指標	$NFI \geq .90$	.86	尚可
		$CFI \geq .90$	.86	尚可
		$TLI \geq .90$	.84	尚可
		$IFI \geq .90$	.86	尚可
	精簡適配指標	$PCFI \geq .50$	.75	佳
		$PNFI \geq .50$	.74	佳
	卡方自由度比	1~3 之間	$\chi^2/df = 19.93$	未盡理想
	所有估計參數	均達顯著水準	是	佳
模式內在品質	個別項目信度	.50	提取訊息	.77 尚可
			概化訊息	.81 佳
			解釋訊息	.89 佳
			整合訊息	.89 佳
	潛在變項組成信度	$\geq .50$	科學多重文本閱讀理解評量	.72 尚可

4. 模式內在結構適配度的檢驗

模式內在結構適配度評鑑結果顯示（參照圖 3），所有估計參數均達顯著水準，一階因素負荷量 λ 介於 .49~.94，分別是 .77、.84、.77、.54、.62、.57、.66、.50、.49、.58、.79、.94、.91、.86、.88、.92、.89、.78、.78。因此，一階因素對指標變項之個別項目信度（Standardized Regression Weights，類似迴歸係數的 R^2 ）依序 .59、.70、.59、.29、.38、.33、.44、.25、.24、.34、.62、.88、.83、.74、.77、.84、.79、.62、.61（因素負荷量的平方），除 S3~S6、S10~S12 外，均 $> .5$ 。四個一階因素被二階共同因素解釋的因素負荷量（即徑路係數 γ ）分別為 .77、.81、.89、.89，是以四個一階因素能被二階共同因素所解釋的變異量，依序為 .60、.66、.80、.80（ γ^2 ），均 $> .50$ 且呈顯著性，亦即本評量之指標信度尚佳，表示 SMTRCA 假設模式的內在品質呈現可接受範圍。

（二）抽取變異比

抽取變異比分析結果顯示，一階因素之成分效度（ $\rho_r = (\sum \lambda^2) / [(\sum \lambda^2) + (\sum \epsilon)]$ ）依序：「提取訊息」為 .63、「概化訊息」為 .32、「解釋訊息」為 .78 與「整合訊息」為 .73；二階因素之組合效度為 .72，除「概化訊息」外，其餘均 $> .50$ ，顯示 SMTRCA 之成分效度與組合效度尚在可接受範圍內。

綜上所述，依據「SMTRCA 的因素效度考驗：二階 CFA 模式」檢驗結果顯示，四因素之 SMTRCA 二階模式與資料之適配度尚稱符合。

討論

一、綜合討論

「科學多重文本閱讀理解評量」所評測的認知能力，包括：(1) 讀者對於科學文本訊息的字義理解（提取訊息）；(2) 讀者對於科學文本訊息的大意理解（概化訊息）；(3) 審視讀者建構服膺全文主旨且具融貫的情境模式（解釋訊息）；(4) 檢視讀者對於科學文本訊息據以進行文本內及文本間的評鑑與批判理解（整合訊息）。茲將虛列研究結果暨綜合討論，並針對本研究結果，提出未來可供學術社群應用，暨研究者仍待深究議題的建議。

（一）SMTRCA 之信度尚在可接受範圍

1. 評分者內的一致性

分析結果顯示，評分者內之 *Cronbach's α* 值均 $> .9$ ，顯示 3 位評分者內的一致性相當穩定。

2. 評分者間的一致性

（1）古典測驗理論的分析

本研究分析結果顯示，4 個向度分評量的 *Kendall ω* 和諧係數均 $> .80$ ，並達顯著相關 ($p < .001$)，表示評分者間的一致性相當好。然而，這種一致性指標（consistency index），意指評分者間有著都高或都低之相對等級的評分趨勢。Eckes（2009）指出高評分者間的信度是解決評分者變異的標準方法，但並不表示評分的正確性，極有可能導致錯誤的結論。事實上，評分者變異會隱蔽評測的構念而影響實作評量的效度與公平，是一個不被期望介入受試者分數的變項。因此本研究繼而輔以現代測驗理論的分析，以確認評分者效應是否干擾評量的精確性，致使造成受試者能力估計的偏誤，藉以確認評量的品質。

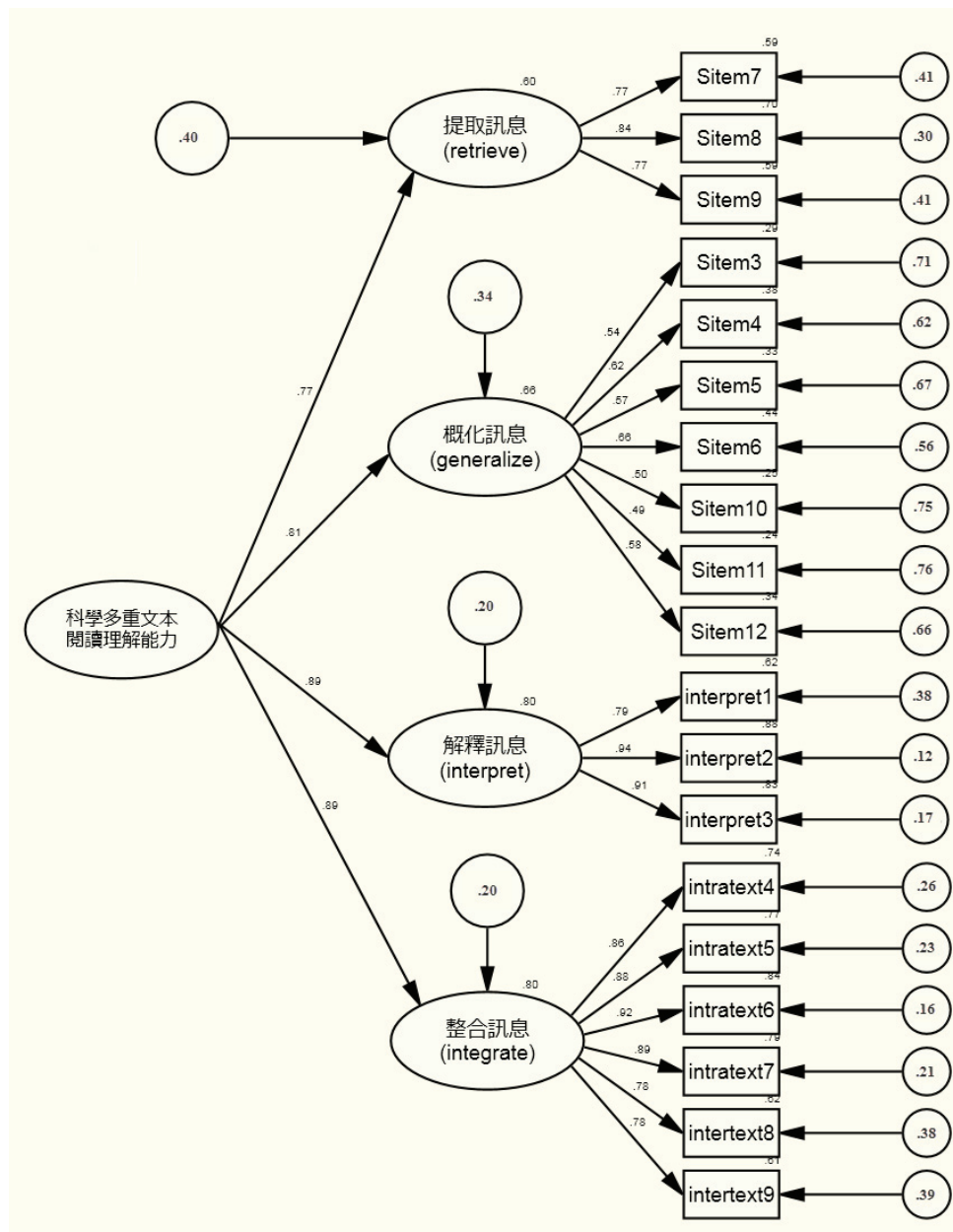


圖 3 SMTRCA 之因素分析結構 (標準化解)

註：+係數係指設定1之參數變項

(2) 現代測驗理論的分析

a. 卡方考驗

從 Rasch 的觀點，申論的評閱，透過多元觀點專家去評閱單一特性構念會有較好的效果。MFRM 分析結果顯示評分者間的嚴苛度確實存在差異，並能有效區分出高、低能力的學生。另 RSM 及

PCM 模式比較的分析結果顯示實徵資料較適配 PCM，表示評分者間的閾值嚴苛度存在差異（參照表 10）。

b. BIC 準則

BIC 適配指標結果顯示實徵資料與 PCM 較為適配，表示評分者間存在不同評分的閾值嚴苛度。

綜上所述，古典測驗理論的分析結果顯示，評分者內與評分者間的一致性相當好。MFRM 分析結果顯示評分者間確實存在嚴苛度的差異（0.082~0.506 logits），符合 Rasch 觀點，即透過多元觀點專家進行單一構念之申論的評閱，會有較好的效果！亦滿足 Rasch 模式中每個評分者具獨特嚴苛度的假設前提。同時與 Rasch 期望模式的適配程度在理想範圍，表示無論是嚴格或者寬鬆的評分者，均能有效區分高、低能力的受試者。此外，透過 RSM 及 PCM 模式比較的卡方考驗，以及 BIC 適配指標，均顯示實徵資料均較適配 PCM，意指評分者閾值嚴苛度存在顯著差異，此結果亦符合一般對於人評分的預期。由於評分者必須植基自身對評量構念的理解，並使用計分標準判斷受試者對該構念試題或任務所創造的反應，這一連串包含詮釋、評估、評分的認知歷程鏈，牽涉到評分者的心理計量，故而很難像機器一樣的評分（machinelike scoring）暨達到像機器般的一致性（machinelike consistency），是可以被理解與接受的。

3. 內部一致性信度

Cronbach's α 分析結果顯示，除「提取訊息」、「概化訊息」外，其餘均 > .70，全評量 α 則在 .90 以上，顯示 SMTRCA 檢具內部一致性之信度。

4. 個別項目信度

個別項目信度分析結果顯示，第一階因素負荷量的平方（ λ^2 ），除「提取訊息」之 S3~S6、S10~S12 外，其餘均 > .50。四個一階因素能被二階共同因素解釋的百分比（ γ^2 ）均 > .50 且呈顯著性，表示本評量之指標信度尚在可接受範圍。

5. 建構信度

SMTRCA 四個一階因素之建構信度均在 .77 以上，顯示本評量之建構信度尚佳。

（二）SMTRCA 之效度尚在可接受範圍內

1. 探索性因素分析

探索性因素分析結果顯示，所抽出的因素與本評量的結構相符合，且試題結構尚稱穩定。19 題因素負荷量皆在 .33 以上，解釋變異量為 46.33%，表示本評量尚具有一致而穩定的結構。

2. 驗證性因素分析

二階驗證性因素分析結果顯示，模式之基本適配度、整體適配度及內在結構適配度尚在可接受範圍。另四個一階因素之成分效度，除「概化訊息」外，其餘均 > .50，表示本評量之抽取變異比尚在可接受範圍。

（三）SMTRCA 為二階四因素心理構念之假設尚契合觀察資料

SMTRCA 可謂一多構念、階層性之能力評量，其一級因素包括「提取訊息」、「概化訊息」、「解釋訊息」、「整合訊息」四個構念，而二級共同因素，則為科學多重文本閱讀理解能力。模式評鑑結果顯示，多數指標均在可接受範圍，假設模式尚且契合觀察資料。然仍有部分指標，諸如 χ^2 、GFI、RMSEA、增值適配指標、卡方自由度比，未達理想門檻，意謂未來仍應持續蒐集資料，進行模式的驗證與修正，俾使假設模式更趨理想適配之臨界值。

（四）RMTRCA 符合多重文本閱讀理解之階層性認知歷程架構

本研究認為科學多重文本閱讀理解從訊息輸入階段之字詞句型的理解（提取訊息）、摘要大意的歸納（概化訊息），到訊息輸出階段之推論隱含意義暨表述重點訊息（解釋訊息）至演繹觀點進行文本的評鑑與分析（整合訊息），係為一個有階層性的認知歷程。其中訊息輸出階段之跨文本整合能力又更高於整合能力、文本內整合能力及解釋能力。透過 MFRM 分析結果顯示，評分標準難度依序為跨文本整合能力（1.57 logits）、整合能力（-0.05 logits）、文本內整合能力（-0.71 logits）、解釋訊息（-0.81 logits），符合本研究提出的認知階層架構。基此，RMTRCA 符合科學多重文本閱讀理解之階層性認知歷程架構。

二、研究應用

本研究建議未來可植基 SMTRCA 架構，延續進行相關的研究，包括開發科學多重文本閱讀理解評量報表系統、探討科學多重文本閱讀理解認知模式、建置科學多重文本閱讀理解評量題庫，以及評分者變異現象的探討。

（一）進行標準設定建立標準參照模式的決斷分數

過去很多研究將評測的反應表現等化到同一量尺上，透過受試者在評量表現的相對位置，決定決斷分數（*cut-off score*），並藉由百分比將其分級，係為植基相對標準之常模參照模式。因此，受試者的表現水準會因為比較群體的不同，而可能被歸類在不同的水準，亦即表現處於變動的狀態，端視所比較群組的程度而定。因此，常模參照所選定的決斷分數若要具有意義，前提必須滿足受試母群相似，以及試題所植基的雙向細目表要相同的假設（Cizek & Bunch, 2007）。為克服常模參照模式所難達到之精確解釋受試者表現水準的限制，並確保不因比較群體的差異而有所不同，本研究建議未來可比照絕對標準，採用標準參照（*criterion-reference*）模式，透過事先訂定學生表現水準暨所具備知識與能力程度之表現標準設定（*standard setting*）（Cizek, 2006; Eckes, 2009），建立決斷分數，以闡釋學生「應知」及「應能」的內容標準暨達到內容標準的水準程度。

（二）科學多重文本閱讀理解評量報表系統

MFRM 可針對評分者對於評量構念等能力表現的觀點進行詳細分析，並能調整受試者表現反應以彌補評分者嚴苛度的差異。因此，本研究建議未來以 MFRM 估計受試者能力，透過依據嚴苛度調整其反應表現的統計控制，致使科學多重文本閱讀理解的評測獨立且不受評分變異的影響。同時依據標準設定所制訂各表現標準陳述（*performance level descriptions*，簡稱 PLDs）與決斷分數，開發評量報表系統，提供具診斷功能之分析式的回饋評述，協助受試者了解科學多重文本閱讀理解的強項及弱項，作為教師閱讀教學與學生閱讀加強的參考依據。

（三）提出科學多重文本閱讀理解認知模式

本研究透過 MFRM 模式進行任務難度（*criteria*）的估計，結果發現以「跨文本整合能力」（1.57 logits）最難，其次依序為「整合能力」（-0.05 logits）、「文本內整合」（-0.71 logits）、「解釋訊息」（-0.81 logits），符合研究者所提出科學多重文本閱讀理解能力係為四個一階因素所抽出二階之共同因素的心智架構。然而本研究使用 MFRM 對任務難度的估計，僅可粗略窺覽多重文本閱讀理解的認知歷程，若要進一步探討，未來應深入釐清影響科學多重文本閱讀理解表現的因素，並提出科學多重文本閱讀理解認知模式，透過結構方程模式（SEM）進行潛在變項間之因果路徑的驗證分析，以提出為實徵資料所驗證支持的理論模式。

（四）評分者變異現象的探討

由於評分者必須植基對構念的理解，依據評分規準判斷受試者對該構念試題或任務的反應表現。這一連串詮釋、評估與評分的歷程，涉及評分者的嚴苛度、解釋和使用評分規準、理解和使用評分量尺類別，以及評分者一致性程度等心理計量的問題。本研究從評分者與計分規準的交互作用結果發現，*rater 3* 依據評分規準的評閱，符合本研究認知階層架構（參照圖 2， $3.4 (0.30 \text{ logits}) > 3.2 (0.09 \text{ logits}) > 3.3 (-0.11 \text{ logits}) > 3.1 (-0.29 \text{ logits})$ ）。但是 *rater 1*（參照圖 2， $1.3 (0.14 \text{ logits}) > 1.1 (0.12 \text{ logits}) > 1.2 (-0.07 \text{ logits}) > 1.4 (-0.19 \text{ logits})$ ）及 *rater 2*（參照圖 2， $2.1 (0.17 \text{ logits}) > 2.2 (-0.02 \text{ logits}) > 2.3 (-0.04 \text{ logits}) > 2.4 (-0.11 \text{ logits})$ ）則不然，因此評分者變異問題亟待後續進一步深入探討研究。

（五）發展科學多重文本閱讀的策略與技能

Adams 與 Lowery（2007）指出涉及學科的閱讀，除了一般認字、理解等閱讀歷程外，更要有學科的先備知識和閱讀技巧的輔助。故而界定科學多重文本閱讀理解的認知模式，可據此發展以語文閱讀為基礎之科學閱讀的策略與技能，裨益學生能從教科書的閱讀歷程中，獲得有效學習策略的能力，達到終身學習的目標。

三、研究建議

(一) 評量題本架構內涵宜持續修訂

本研究建議未來仍可繼續驗證 SMTRCA 的信度與效度，首先可進行效標關聯效度的分析，以建立更好的內容效度。其次，可擴大受試族群，促使建立標準參照測驗模式，提供使用者更完備的回饋訊息參酌。

(二) 建置科學多重文本閱讀理解評量題庫

有鑑於本研究所發展之 SMTRCA 係屬地球科學議題，為豐富科學多重文本素材，本研究建議未來宜繼續發展涵蓋包括物理、化學、生物等科學題本，藉以建置科學多重文本閱讀理解評量題庫。

(三) 運用發展更成熟之 SMTRCA 進行相關研究

本研究可謂工具編制初探，未來可持續注摺研究以發展更成熟之 SMTRCA，藉以蒐集實徵資料進行相關研究，包括學生多重文本閱讀理解的認知模式、學生理解文本能力及闡釋文本能力的發展階段，甚至探討地域、文化、年級與性別變項，是否會造成學生多重文本閱讀理解能力發展的差異。

參考文獻

- 唐淑華、蔡孟寧、林烘煜 (2015)。多文本課外閱讀對增進國中學生理解歷史主題之研究：以「外侮」主題為例。*教育科學研究期刊*, 60(3), 63-94。[Tang, S. H., Tsai, M. N., & Lin, H. Y. (2015). Using multiple texts to foster Junior High Students' historical understanding. *Journal of Research in Education Sciences*, 60(3), 63-94.]
- 楊蕙如、陸怡琮 (2010)。閱讀目的對成人閱讀多重文本的理解表徵建構之影響。中正大學主辦「台灣心理學會第 49 屆年會」(嘉義)。[Yung, H., R., & Lu, I. C. (2010). *The effects of reading purpose on the representation and comprehension of multiple texts reading of college students*, 49th Annual Meeting of the Taiwanese Psychological Association, Symposium conducted at National Chung Cheng University, Chiayi.]
- Adams, T. L., & Lowery, R. M. (2007). An analysis of children's strategies for reading mathematics. *Reading and Writing Quarterly*, 23(2), 161-177.
- Bennett, R. E., & Ward, W. C. (1993). *Construction versus choice in cognitive measurement: Issues in constructed response, performance testing, and portfolio assessment*. Hillsdale, NJ: Lawrence Erlbaum Associates.
- Bråten, I., & Strømsø, H. (2010). When law students read multiple documents about global warming: Examining the role of topic-specific beliefs about the nature of knowledge and knowing. *Instructional Science*, 38(6), 635-657. DOI: 10.1007/s11251-008-9091-4

- Bråten, I., Strømsø, H. I., & Britt, M. A. (2009). Trust matters: Examining the role of source evaluation in students' construction of meaning within and across multiple texts. *Reading Research Quarterly*, 44(1), 6-28. DOI: 10.1598/RRQ.44.1.1
- Cizek, G. J. (2006). Standard setting. In S. M. Downing & T. M. Haladyna (Eds.), *Handbook of test development* (pp. 225-258). Mahwah, NJ: Lawrence Erlbaum Associates.
- Cizek, G. J., & Bunch, M. B. (2007). *Standard setting: A guide to establishing and evaluating performance standards on tests*. Thousand Oaks: Sage.
- Eckes, T. (2005). Examining rater effects in TestDaF writing and speaking performance assessments: A many-facet Rasch analysis. *Language Assessment Quarterly*, 2(3), 197-221. DOI: 10.1207/s15434311laq0203_2
- Eckes, T. (2009). Many-facet Rasch measurement. In S. Takala (Ed.), *Reference supplement to the manual for relating language examinations to the Common European Framework of Reference for languages: Learning, teaching, assessment* (Section H., pp. 1-52). Strasbourg, France: Council of Europe/ Language Policy Division.
- Hartman, D. K., & Allison, J. (1996). Promoting inquiry-oriented discussions using multiple texts. In L. B. Gambrell & J. F. Almasi (Eds.), *Lively discussions!: Fostering engaged reading* (pp. 106-133). Newark, DE: International Reading Association.
- Johnson, H. M., & Seifert, C. M. (1999). Modifying mental representations: Comprehending corrections. In H. van Oostendorp & S. Goldman (Eds.), *The construction of mental representations during reading* (pp. 303-318). Mahwah, NJ: Lawrence Erlbaum Associates.
- Kintsch, W. (1988). The role of knowledge in discourse comprehension: A construction-integration model. *Psychological Review*, 95(2), 163-182.
- Landy, F. J., & Farr, J. L. (1983). *The measurement of work performance: Methods, theory, and applications*. New York, NY: Academic Press.
- Linacre, J. M. (1989). *Many-facet Rasch measurement*. Chicago, IL: MESA Press.
- Perfetti, C. A., Rouet, J. F., & Britt, M. A. (1999). Toward a theory of documents representation. In H. vanOostendorp & S. R. Goldman (Eds.), *The construction of mental representations during reading* (pp. 99-122). Mahwah, NJ: Erlbaum.
- Rouet, J. F. (2006). The skills of document use: *From text comprehension to web-based learning*. Mahwah, NJ: Lawrence Erlbaum Associates.
- Rouet, J. F., Britt, M. A., Mason, R. A., & Perfetti, C. A. (1996). Using multiple sources of evidence to reason about history. *Journal of Educational Psychology*, 88(3), 478-493. DOI: 10.1037/0022-0663.88.3.478

- Rouet, J. F., Vidal-Abarca, E., Erbou, A. B., & Millogo, V. (2001). Effects of information search tasks on the comprehension of instructional text. *Discourse Processes*, 31(2), 163-186. DOI: 10.1207/S15326950DP3102_03
- Schwarz, B. (2003). Collective reading of multiple texts in argumentative activities. *International Journal of Educational Research*, 39(1-2), 133-151. DOI: 10.1016/S0883-0355(03)00077-6
- Valenti, S., Neri, F., & Cucchiarelli, A. (2003). An overview of current research on automated essay grading. *Journal of Information Technology Education*, 2(1), 319-330.
- van den Broek, P. (1990). The causal inference maker: Towards a process model of inference generation in text comprehension. In D. A. Balota, G. B. F. d'Arcais, & K. Rayner (Eds.), *Comprehension processes in reading* (pp. 423-445). Hillsdale, NJ: Lawrence Erlbaum Associates.
- Wolfe, M. B., & Goldman, S. R. (2005). Relations between adolescents' text processing and reasoning. *Cognition and Instruction*, 23(4), 467-502. DOI: 10.1207/s1532690xci2304_2

收 稿 日 期：2016 年 03 月 17 日

一稿修訂日期：2016 年 08 月 03 日

二稿修訂日期：2016 年 10 月 31 日

三稿修訂日期：2016 年 12 月 15 日

四稿修訂日期：2017 年 01 月 25 日

五稿修訂日期：2017 年 02 月 02 日

接受刊登日期：2016 年 12 月 26 日

Bulletin of Educational Psychology, 2017, 49(2), 215-241

National Taiwan Normal University, Taipei, Taiwan, R.O.C.

Developing and Validating a Scientific Multi-Text Reading Comprehension Assessment: Evidence from Texts Describing Relationships between Climate Changes and the Three Gorges Dam

Hsiao-Hui Lin

Graduate Institute of Science Education
National Taiwan Normal University

Yu-Htsuen Tzeng

Center for Teacher Education & Institute of
Education National Chung Cheng University

This study aimed to advance the Scientific Multi-Text Reading Comprehension Assessment (SMTRCA), with a focus on the Rubric of Multi-Text Reading Comprehension Assessment (RMTRCA) designed to evaluate the extent of reading comprehension. To this end, we used scientific texts describing the dispute of the relationships between climate changes and the Three Gorges Dam and developed assessment items according to our rubric. Test items included 10 close-ended and 9 open-ended questions and were categorized into 4 subscales: information retrieval, information generalization, information interpretation, and information integration. The results of analysis showed that the cronbach's α values were more than .9, indicating that the intra-rater consistency was well. Secondly, the Kendall's coefficient of concordance was more than .8 and its P value was smaller than .001, denoting a consistent scoring pattern between raters. Additionally, the analysis of many-facet Rasch measurement (MFRM) and the comparison of the rating scale model (RSM) and the partial credit model (PCM) showed that the chi-square test of rater severity and threshold difficulty were significant. The infit and outfit MNSQ of the former are between 1 ± 0.3 , meaning that both severe and lenient raters can distinguish high-ability students from low-ability students more effectively. The latter means that the rating procedures involve human interpretation, evaluation and scoring processes so that it is difficult to reach a machine-like consistency level. However, this is in line with expectations of typical human judgment processes. Thirdly, most values of Cronbach's α of test items were larger than .7 except those from information retrieval and information generalization but overall they were all within acceptable range. Finally, confirmatory factor analysis showed that there was an acceptable goodness-of-fit among the SMTRCA. The SMTRCA accounts for .60, .66, .80, and .80 of the variance associated with the first order factors of 4 subscales.

KEY WORDS: Confirmatory factor analysis, Many-facet Rasch measurement, Partial credit model, Rating scale model, Scientific multiple text reading comprehension assessment

