創新電腦化測驗題型與測量模式的結合

陳柏熹
國立臺灣師範大學教育心理與輔導學系副教授

摘要

隨著電腦科技的進步，測驗的題型、作答方式與試題媒材也有明顯的改變。從傳統的紙本劃記或書寫型式的測驗題型，擴展到應用資訊設備來點選答案、圖形圈選、拖曳、搖桿操作、電腦繪圖，甚至於表現動作再以攝影機記錄的方式來作答。然而，應用電腦科技於測驗中的主要目的是為了提升測驗的信度與效度，因此，是否達到提升測驗信度與效度的目的才是評價創新電腦化題型的關鍵。本文的觀點是將電腦化測驗題型與測量模式結合，讓創新的電腦化測驗題型有驗證效度與評估信度的理論基礎，並以兩項測驗為例來闡述如何將創新電腦化測驗題型與測量模式作結合。

關鍵字：電腦化測驗、測量模式、試題反應理論、信度、效度

一、創新電腦化測驗題型

電腦科技在近二十年來有相當快速的發展，相關的資訊技術也廣泛地應用到當代的許多心理與教育測驗中。舉凡大型教育測驗中的托福（Test of English as a Foreign Language, TOEFL）與 GRE（Graduate Record Examinations）考試、國際教育評比 PISA（the Programme for International Student Assessment）、語言能力認證中的漢語水平考試（Hanyu Shuiping Kaoshi, HSK）與華語文能力檢定（Test of Chinese as a Foreign Language, TOCFL），以及許多職業證照考試，例如，美國醫師證照考試（United States Medical Licensing Examination, USMLE）、護士證照考試（National Council Licensure Examination for Registered Nurse, NCLEX）、會計師、建築師證照考試…等，不勝枚舉，都是採用電腦化測驗題型。除了上述的成就測驗（achievement test）之外，性向測驗、性格測驗與職業興趣量表也都開始有電腦化版本，例如，著名的通用性向測驗（General Aptitude Test Battery, CAT-GATB）與大五人格量表（Revised NEO Personality Inventory, NEO PI-R）都有電腦化適性測驗或電腦化施測與解釋版本。其中，美國陸軍職業性向電腦化適性測驗（Armed Services Vocational Aptitude Battery, CAT-ASVAB）可以算是最早發展出電腦化適性測驗（computerized adaptive testing, CAT）且研究較完整的電腦化適性測驗。

然而，上述測驗的題型大部分是選擇反應題型（select response item），例如，選擇題（multiple choice item）或李克特量表（Likert type scale），其電腦化版本只是將此題型從紙本形式變成電腦作答而已，並沒有利用電腦多媒體技術來提高測驗的效度或信度。拜現代電腦科技之賜，測驗除了可以使用這兩種題型，還

有許多不同類型的作答方式與題目呈現形式。Parshall、Davey 與 Pashley（2000）以及 Parshall 與 Harmes（2007）曾提出當代電腦測驗題型的多向度分類架構圖（taxonomy of innovative item），將電腦化測驗題型從評量結構（assessment structure）、複雜度（complexity）、似真性（Fidelity）、互動程度（interactivity）、媒體類型（media inclusion）、作答反應方式（response action）、計分方式（scoring method）等七個向度來進行分類，以下分別簡述之：

（一）評量結構

評量結構是指題目之間的結構關係，通常可以依此將題目分成獨立題（discrete item）、情境任務題組（situated task）、虛擬實境活動（simulated environments）。獨立題是傳統單一題目的問題，又可以分為選擇反應題（selected response item）與建構式反應題（constructed response item）兩大類。選擇反應題型如是非題、選擇題或配合題等，其具有作答快速與計分客觀的優點，但比較容易讓受測者產生猜答案的行為而影響測量的信度。建構反應題型如填空題、簡答題、計算題、申論題等。由於受測者難以在這種題型中猜測答案，因此可以改善選擇反應題型容易被猜對的誤差，但作答較費時，且計分也比較容易因為評分者的主觀因素影響評分者一致性信度。圖 1 是多元智慧電腦化適性測驗的題目（陳柏熹，2008），屬於選擇反應題型，受測者需分別針對每個題目點選答案。圖 2 為節能減碳素養電腦化測驗系統（Chen, Huang, Liu & Chen, 2014）的建構反應試題，受測者需打字回答問題，雖然是採用題組式（testlet）題型，但各題目間仍是獨立試題

圖 1 多元智慧電腦化適性測驗之選擇反應題

圖 2 節能減碳素養電腦化測驗之建構反應題

情境式任務是一系列具有關連的情節式（scenario）測驗題組，受測者需要多個步驟才能完成題目所要求的任務，如圖 3 所示。在這類題目中，各子題之間可以是具有結構式（structured）的路徑關係，其前項任務是下一個任務的基礎，也可以是非結構式（structured）的開放型任務，由受測者自行決定如何進行。圖 3 中小學資訊能力評量系統（張國恩、劉遠禎、王曉瑋、蕭顯勝、宋曜廷、陳柏熹，2008）的情境式任務題組就是屬於結構式的情境任務，受測者需要依據題目的要求逐一進行各項任務。圖 4 的電腦化創造力測驗（陳柏熹，2011）則是屬於非結構的開放式任務，受測者可以根據題目的要求，自行決定要如何進行操作，設計出一項具有創新性與實用性的家具，以達成評量目標。

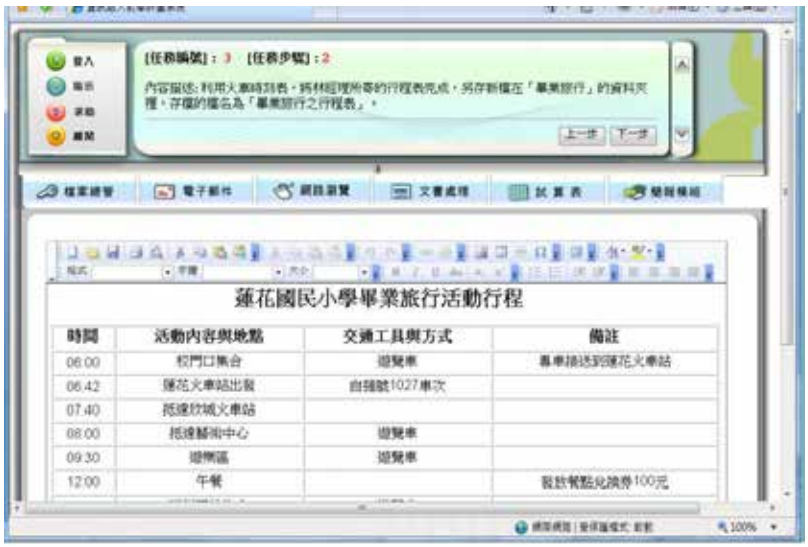

圖 3 中小學資訊能力評量系統中的結構式情境任務題型
（圖片來源：中小學資訊能力評量機制發展與推廣計畫成果報告）

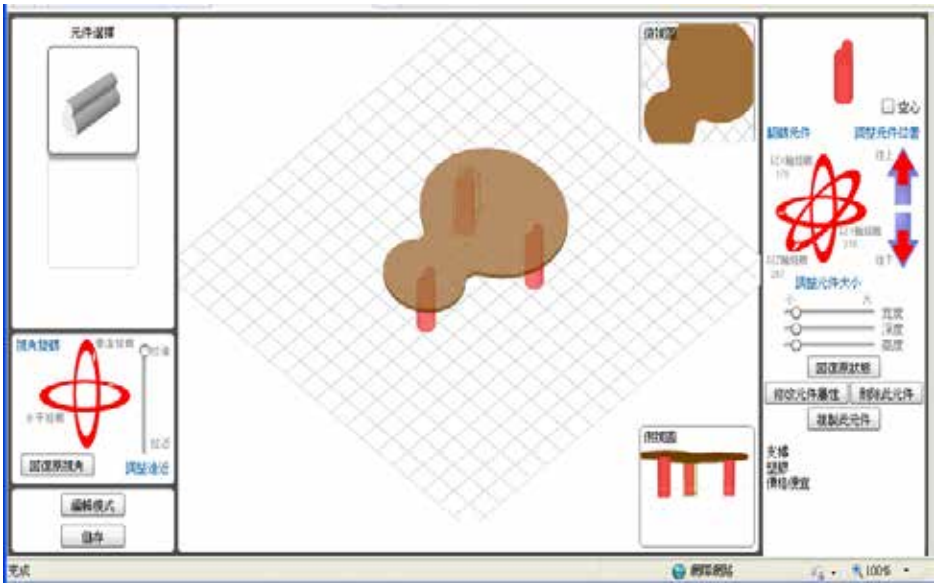

圖 4 電腦化創造力測驗中的非結構式情境任務題型

虛擬實境活動是讓受測者在模擬真實狀況的電腦虛擬實境測驗系統中進行實作活動，例如：飛機駕駛、汽車、儀器操作，如圖 5 所示。受測者必須進行實際的操作，就好像在真實的生活環境中進行實務操作一般。

圖 5 汽車駕駛能力（左）與手眼協調性（右）的虛擬實境評量題型
（左圖取自林正堅（2011）；右圖取自蕭顯勝、林建佑、邱敬尊（2011））

（二）複雜度

題目複雜度是指受測者在回答題目時所需考慮的訊息種類與數量多寡。例如：在題目中的不同位置呈現多項文字與圖形，讓受測者進行判斷。還有些題目需使用圖形組合、按鈕操作、表單輸入或數值計算、影片播放等複雜的電腦操作活動等。題目的複雜度愈高，認知負荷量就愈大。傳統僅包含一段文字或一張圖形的選擇題是屬於低複雜度的題目，而情境任務題組通常是較高複雜度的題目。有些國家的醫學證照考試採用電腦化的模擬個案臨床測驗，就是屬於

高複雜度的測驗，受測者需要判斷許多不同的訊息內容，包含病人的主訴與外觀症狀、醫學檢驗報告、用藥紀錄等，針對這些訊息進行診斷或後續檢查，並開藥物或安排住院治療。例如：美國醫師證照考試電腦化模擬個案測驗（United States Medical Licensing Examination step 3 computer-based cases simulation, USMLE step 3 CCS）（USMLE, 2014）。而圖 4 的電腦化創造力測驗也是屬於高複雜度的題型，該測驗中有許多的元件操作方式與調整按鈕，受測者需要了解這些元件如何使用，才能有效地完成題目的任務。

（三）似真性

似真性是指題目內容與真實生活環境的相似程度，題目呈現出愈多與真實生活情境相似的題材內容，或是要求受測者作出與生活情境相似的實作活動，題目的似真性就愈高。傳統以文字為基礎的選擇題其似真性較低，而虛擬實境實作活動是屬於高似真性的題目。題目的似真性愈高，表示愈接近生活情境中的能力展現，測量的效度就愈高。以圖 5 中的汽車駕駛測驗為例，若是全部採用文字型選擇題，只能測量出受測者的汽車駕駛常識或交通規則，即使受測者在該測驗中得到高分也不能保證他已經會開車了。若使用圖 5 中的高似真性汽車駕駛虛擬實境測驗，則較能測量出受測者在面對道路駕駛情境時操作方向盤與控制車輛的能力及應變技巧。若該題目情境與使用的設備材料似真性愈高時，所測量到的能力就愈接近真實的汽車駕駛能力。

（四）互動程度

互動程度是指根據作答者的反應，即時呈現出與此反應有關的資料、情境或後續題目，讓受測者進行下一階段的答題。傳統測驗的互動程度較低，通常只呈現題目要求受測者回答答案就結束了。而互動性較高的測驗會要求受測者回答一連串有邏輯關連的問題，不宜將這些題目拆開使用。在美國的醫師證照考試的模擬個案測驗中，題目會先提供有關病人的症狀與檢查數據，接著會讓受測者進行診斷、用藥、進行後續追蹤檢查等一系列問題，且受測者回答完每個問題後，會得到回答後的結果，前項處理的結果就是進行後續處理的依據（USMLE step 3 CCS; USMLE, 2014）。在 Liu, Lin 與 Wang（2012）的統計概念模擬輔助學習系統中，受測者可以藉由操作資料數據立即看到統計圖表上的變化，並藉此來進行後續的概念學習與判斷作業，也是屬於高互動性的學習評量系統。

互動式測驗的試題資訊會隨著受測者的反應而改變，因此計分方式相當複雜，目前此類測驗尚未找到適合的心理計量模式。較常採用的作法是將受測者的作答反應與專家群所建立的最佳反應（optimal response）進行比對，並分別從多個向度來計分。以美國的醫師證照電腦化模擬個案測驗為例，就是根據專家群所建立的最佳反應，將受測者的作答反應依資訊考量的完整性、時效性、治療的有效性、避免風險等多個向度來進行評分。

（五）使用的媒體類型

腦化測驗中常見的題目媒體類型主要有文字、圖形或照片、影片、聲音、動畫等。媒體的使用通常與題目所欲測量的能力有關。例如：在溝通能力測驗中，受測者在進行溝通時可能需要面對訊息傳遞者或訊息接受者，因此可以在題目中預先錄製一段影片（或動畫），並要求受測者在看完這段影片後去判斷影片中的主角所要傳遞的訊息、或是同理影片中主角的感受、或是回答出如何說服他人或安慰他人等表達語句。在圖 6 的電腦化大學生基本素養測驗（陳柏熹，2014）中，其溝通表達素養就是要求受測者在看完一段動畫後，進行上述的判斷。此外，在該測驗的美感素養試題中，由於主要是評量受測者對美感訊息的理解與表達能力，因此大部分題目都是以生活中的照片、圖案或影片為媒體來呈現，並要求受測進行判斷。

溝通表達素養試題　　美感素養試題

圖 6 電腦化大學生基本素養測驗

（六）作答方式

由於當代電腦科技的快速發展，目前受測者在電腦上作答的方式相當多，包括用滑鼠點選、螢幕觸控、鍵盤輸入數值或文字、以搖桿控制、在圖形中圈選位置（hot spot）、圖案拖曳或組合排列、用麥克風錄音等。早期電腦化測驗的作答方式通常是以鍵盤輸入或滑鼠點選，這類測驗大多是以認知能力測驗為主。近代則可以用攝影機來擷取受測者的肢體動作反應來進行作答，或甚至於用眼動儀追蹤眼球運動，以眼睛控制的方式來作答。這種較新的作答技術通常被應用在實作技能的評量上。圖 7 是以動作影像擷取作答方式來進行肢體動作智慧的電腦化測驗（陳柏熹、歐詠芝，2011b）。在此測驗中，題目刺激與受測者的身體會呈現在電腦螢幕上，受測者必須根據題目的要求來移動身體，使螢幕中的身體影像能碰觸到特定圖案或避免碰觸到某個外框，以評量出受測者的反應速率與平衡感。

反應速率測驗　　平衡感測驗

圖 7 肢體動作電腦化測驗的動作影像擷取作答方式

（七）計分方法

計分方式是指如何根據受測者的作答反應計算出每一題的原始分數。獨立題的選擇反應題型計分方式較單純，通常是採用二元計分（dichotomous scoring）。而獨立題的建構反應題型、情境任務題組或虛擬實境活動等，由於評量結構與作答方式都較複雜，通常是多元計分（polytomous scoring）。

二、創新電腦化測驗題型對測驗效能與測量模式的影響

表 1 是電腦化測驗題型的分類向度對測驗效能與測量模式的主要影響層面。雖然每一種分類方式中的不同題型都可能會同時影響到測驗的信度、效度與分析時所採用的測量模式，但有些分類向度對測量模式的影響較大，有些則是對測驗效度的影響較大。分述如下：

題目的評量結構類型主要與所使用的測量模式有關。例如：在獨立題中，每個題目彼此是互相獨立的，因此適合使用試題反應理論（item response theory, IRT），因 IRT 中的基本假設是局部獨立性，也就是對相同能力水準的受測者而言，各題目的答對機率是互相獨立的。然而，在情境任務題型中，前面題目可能會影響後面題目的答題表現，已經違反局部獨立性假設，此時就須要採用題組反應模式（testlet response model）（Wainer, Bradlow, & Du, 2000; Wang & Wilson, 2005），或是將各個步驟的作答結果與成績合併成一題，改採部份給分模式（partial credit miodel）（Master, 1982）來進行分析。若是使用非結構式情境任務題或虛擬實境題型，由於需要有評分者來評量受測者的實作技能表現，可能會受到評分者主觀因素的影響，因此可以採用多面向模式（many-facet model）（Linacre & Wright, 1993），將評分者嚴苛度與不同實作技能面向都納入測量模式中進行分析。

題目的複雜度也與測量模式有密切關聯。在低複雜度的情況下，影響作答反應的能力向度較單純，因此適合以單向度

測量模式（單向度 IRT 或單一因素的因素分析）進行分析。當題目的複雜度增加時，所包含的訊息種類與數量也增加了，影響作答反應的能力向度或因素也隨之增加，因此較適合採用多向度試題反應模 式（Mckinley & Reckase, 1983; Hattie, 1981; Adams, Wilson & Wang, 1997）或多因素的因素分析來進行分析，如此才能瞭解不同因素對作答結果的影響。除此之外，若將題目所包含的各類訊息視為影響題目難易度的預測變項，也能採用線性對數模式（linear logistic test model, LLTM）（Fischer, 1973）來進行分析，以了解不同訊息對題目難易度的影響力。

似真性主要是影響測驗的效度。當題目情境與使用的設備材料似真性愈高時，所測量到的能力就愈接近真實生活中所展現出來的能力。因此，題目的似真性是決定測驗預測效度（predict validity）的重要因素。

題目的互動程度則是對測驗的信度與效度都會產生影響。由於目前尚未找到適合的心理計量模式可對高互動程度的題目進行分析，且仰賴專家群所建立的最佳反應來與受測者反應進行比對的做法，其客觀性與有效性也缺乏實徵研究證據。因此，目前這種因素對測驗效能的影響力還需進一步研究來提供證據。

媒體類型主要會影響測量的效度。題目中的媒體類型與真實生活愈接近，題目就愈能反映出受測者在日常生活中的情境，其所引發的作答反應就會愈接近受測者的真實能力。不過，閱讀或觀看媒體通常會增加作答時間，因此當測驗的作答時間限制較嚴格時，複雜媒體的題型可能會影響到測量的信度。

電腦化測驗的作答方式必須配合測驗目標，才能有效測量出受測者的能力，提升測驗的效度。例如：若要測量受測者對機械手臂的操作能力，應先了解實際生活狀況中的機械手臂通常是以何種方式來控制？若是以搖桿來控制，則搖桿將會是該項電腦化測驗中較佳的作答方式。圖 5 與圖 6 的幾種作答形式都是實際生活中展現出該測驗目標能力時最常使用的方式。

計分方式與測量模式有密切的關係。若題目為二元計分，則較適合採用二元計分的試題反應模式來進行分析，例如：Rasch 模式（Rasch, 1986）、2PL 模式或 3PL 模 式（Birnbaum, 1968; Lord, 1952）等。當題目為多元計分時，則較適合採用評定量尺模式（rating scale model, RSM ）（Andrich, 1978）或部份給分模式（partial credit model, PCM）（Masters, 1982）。

表 1　電腦化試題分類向度對測驗效能與測量模式的影響

	主要影響測驗的層面		
	測量模式	測驗信度	測驗效度
電腦化試題分類向度			
評量結構	√		
複雜度	√		√
似真性			√
互動程度		√	√
媒體類型			√
作答反應方式			√
計分方式	√		

三、電腦化測驗題型與測量模式的結合

為了讓創新電腦化測驗題型能發揮其測量功能，電腦化測驗編製者需思考如何選擇適當的題型，以及如何選擇適當的測量模式來進行分析。以下以兩項電腦化測驗的研發為例來進一步闡述。

（一）電腦化大學生基本素養測驗

1. 測驗目標

該測驗的目的是評量大學生在創新領導、問題解決、終身學習、溝通合作、公民素養、美感素養、科學思辨、資訊素養、生涯發展等九項通識教育基本素養上的學習成果，屬於生活基本能力的素養（Chen, Hsu, Huang, Li, Chen, Yang & Yeh, 2013）。測驗中每一種素養皆包含認知能力與情意態度兩個面向，本文僅分析認知能力層面之試題。

2. 選擇電腦化測驗題型

有關電腦化大學生基本素養測驗的題型選擇整理如表 2 所示。在評量結構方面，由於該測驗的目標特質與日常生活基本素養有關，因此需要有　日常生活的情境」做為試題題材，引導學生展現出與生活情境有關的基本素養，因此適合採用情境式任務題組。但由於評量目標為認知能力，而非實作技能，因此將任務導向的實作題型改成比較適合評量認知能力的非任務型情境測驗題組。又為了能在學生作答完後即時回饋測驗結果，因此採用可以進行電腦自動計分的選擇反應題型，包括選擇題與叢集式是非題（cluster true-false item）兩種題型，如圖 6 所示。

在複雜度方面，該測驗中各素養欲評量的能力皆為單向度能力，為了避免受到其他能力的干擾，因此選擇低複雜度的題型。而在試題似真性方面，為了使測量結果較能反映出學生在日常生活中的實際能力，因此選擇中等似真性的題型，將生活中常看到的影像、照片或圖表都納入作為題目的材料。但礙於經費有限，因此並沒有發展成高似真性的虛擬實境題型。在互動程度方面，為了讓能力估計單純化，提高計分的客觀性，因此採用低互動性的試題，每個題目在作答完後並不會根據測者的答案而變化後續題目所提供的資訊。

在媒體類型方面，為了提高測驗的效度，該測驗配合各素養的需求而使用各種不同媒材的試題，例如：溝通合作素養多採用人際互動影片或動畫作為題材，美感素養多採用生活中常見的圖像或影片作為題材，問題解決素養多採用圖表呈現問題讓受測者進行問題解決。在作答方式上，由於該測驗是由大學生自行上網作答，且所有試題皆為選擇反應題型，因此該測驗使用大學生最熟悉的滑鼠點選方式，以避免受測者對作答方式不熟悉而影響測量結果。

關於計分方式，由於該測驗中有些素養的題目並無唯一的正確答案，而僅有較佳選項、次佳選項與不佳選項之區別，因此不宜全部採用二元計分模式，需使用多元計分，例如較佳選項記為 2 分，次佳選項記為 1 分，不佳選項記為 0 分。

表 2　電腦化大學生基本素養測驗在電腦化試題分類向度上的題型選擇

電腦化試題分類向度	
評量結構	情境測驗題組
複雜度	低
似真性	中
互動程度	低
媒體類型	多樣
作答反應方式	滑鼠點選
計分方式	多元計分

3. 選擇測量模式

為了能符合上述的情境測驗題組結構與多元計分方式，該測驗採用多元計分的題組反應模式進行分析（Wang & Wilson, 2005）。資料來源為 2012 年 11 月至 2013 年 6 月該測驗進行預試時所蒐集之 1049 位大學生作答反應，該預試採用不等組變動共同題設計（Non-equivalent group with variable anchor test, NEVAT）（Chen, Kuo & Sung, 2011）。表 3 為大學生基本素養測驗之多元計分的題組反應模式分析結果，表中數值為主要欲評量的能力質變異數與題組效果變異數，其中，題組效果變異數愈大，表示題目間的相關係愈高，對真實能力估計的影響就愈大。在估計能力時若未採用題組模式將題組效果分離出來，將會錯估受測者能力值並高估能力估計的精準度（Wang & Wilson, 2005）。從表中可以看出，公民素養、美感素養與資訊素養中皆有部分題組有較高的題組效果。究其原因為該題組中皆有叢集是非題。若將叢集是非題重新計分為一題多元計分題，則題組效果明顯下降。表 3 最後一列是將資訊素養的叢集是非題重新計分後的題組效果分析結果。

表 3　大學生基本素養測驗之多元計分的題組反應模式分析結果

基本素養名稱	能力 σ^2	題組 1 $\sigma^2_{\gamma_{1Y:Y}}$	題組 2 $\sigma^2_{\gamma_{2Y:Y}}$	題組 3 $\sigma^2_{\gamma_{3Y:Y}}$	題組 4 $\sigma^2_{\gamma_{4Y:Y}}$
溝通合作	0.21	0.35	0.91	0.86	-
科學思辨	0.27	0.66	0.56	0.42	-
公民素養	0.27	0.41	0.45	1.14	-
終身學習	0.47	0.59	0.46	0.84	0.92
美感素養	0.40	0.87	0.31	1.17	-
創新領導	0.36	0.36	0.92	0.92	-
問題解決	0.19	0.29	0.64	0.39	-
問題解決	0.19	0.29	0.64	0.39	-
生涯發展	0.37	0.36	0.75	0.64	-
資訊素養	0.15	0.08	0.05	1.12	3.42
資訊素養（recode）	0.17	0.18	0.10	0.33	0.98

（二）電腦化創造力測驗

1. 測驗目標

該測驗主要是評量學生在產品設計的創造歷程與創造成品中所展現出來的創新性與實用性（陳柏熹，2011a）。創造歷程之評分有七項指標，包含功能聯結（含元件選取、材質選取）、表徵轉換（材質用途創新、元件變化）、心向旋轉（元件翻轉）、組合（組合獨創、組合多樣）四大面向。創造成品之評分有八項指標，分別為外觀創新、功能創新、不對稱性、精緻化、多元功能、平衡感、使用便利性、堅固耐用性等。

2. 選擇電腦化測驗題型

電腦化創造力測驗的題型選擇整理如表四所示。在評量結構方面，該測驗的目標特質是設計產品的創造力，為了能適用於不同年齡受測者，所以使用一般人在日常生活中都會接觸到的　家具」設計做為試題題材，因此也是適合採用情境式任務題組。而其任務是使用三項幾何元件材料，實際設計出一項具有創新性及實用性的家具，因此是屬於實作技能類的任務。如表 4 所示。

在複雜度方面，由於創造力為複雜的多向度能力，在產品設計領域被定義為設計出具有創新性與實用性的產品，並被群體所認可（葉玉珠，2004; Amabile, 1996），且可以展現在創造的歷程中，也能展現在創造出來的成品中，因此選擇高複雜度的題型，給予許多控制元件讓受測者可以從不同角度、調整物件的外觀形狀材質，並組合各項物件。

在試題似真性方面，為了使測量結果較能反映出學生在產品設計的實際能力，因此選擇高似真性的題型，將施測介面規劃成與市售產品設計軟體類似的介面，以模擬產品設計人員的工作。在互動程度方面，為了能紀錄受測者的產品設計歷程的各項動作，因此採用高互動性的試題，使受測者在進行每一項操作時都能立即得到回饋，以便調整後續的操作。

在媒體類型方面，由於本測驗所使用的設計元件都是幾何元件，較為單純，因此媒體類型都是使用圖形。在作答方式上，由於該測驗主要目的是要求受測者根據所提供的幾何圖形元件進行移動、變化、組合等操作，因此採用滑鼠點選、圖形拖曳、圖形組合等作答方式。此外，為了使評分者了解作品特性與創作理念，在設計結束時也讓受測者加入文字敘述。

關於計分方式，由於創造力並無唯一的正確答案，而是要從七項創造歷程指標與八項創造成品指標來評估受測者的創造力，因此使用多元計分，每項指標皆為 0-2 分，評分規準參見陳柏熹（2011a）。

表 4　電腦化大學生基本素養測驗在電腦化試題分類向度上的題型選擇

電腦化試題分類向度	
評量結構	開放式情境測驗題
複雜度	高
似真性	高
互動程度	高
媒體類型	圖形
作答反應方式	滑鼠點選、圖形拖曳、圖形組合、文字輸入
計分方式	多元計分

3. 選擇測量模式

為了能符合上述的高複雜度、多種作答反應及多元計分方式，該測驗採用多向度能力之多元計分模式來進行分析，並使用結構方程模式（structure equation model）（Jöreskog & Sörbom, 1993） 來探討模式與資料的吻合程度，受測者為217 位大學生。初始模式採用測量目標中的評分指標架構，其參數分析結果如圖 7 所示。其中 * 表示該項參數之 t 檢定值顯著不等於零（p<0.05）。從創造歷程評分的測量模式中可以看出，「旋轉翻轉」項目信度頗低（0.03=1-0.97），且受潛在變項之影響也未達顯著水準（λ=0.20，p>0.05），表示該指標測量效果不佳，無法達到應有之測量功能，因此該指標在後續分析中予以刪除。而「組合多樣性」雖然受組合操弄潛在變項之影響也不大（λ=0.25，p>0.05），但根據修改指標顯示其在「表徵轉換」潛在變項中之測量效果頗佳。經研究團隊討論其評分規準發現該項評分內涵與「表徵轉換」有關，因此保留該指標併入表徵轉換的潛在變項，改稱為「轉換變化」。創造成品評分測量模式分析結果顯示，各項評分指標之 λ 值皆大於 0.5，且 t 檢定皆顯著不等於零（p<0.05），顯示成品評分向度之測量模式尚可。但根據修改指標顯示，「外觀創新」與「不對稱性」、「功能創新」與「多元功能」兩組指標皆有顯著之相關，顯示此兩組指標在評分時可能會互相干擾，經研究團隊討論發現此兩組指標內的兩項指標之內涵頗為接近，因此決定兩組指標中都只保留一項指標，刪除「不對稱性」與「多元功能」。

從結構關係模式來看，潛在自變項與潛在依變項之參數值 γ 皆不高，且皆未達顯著。但由於刪除其中某些 γ 參數時，會影響到其他 γ 參數之估計。因此，本研究之後續分析僅將參數絕對值 <0.2 之 γ 先行刪除，其他 γ 仍予以保留。

初始模式之模式符合度資料如表五所示。從表五中可以看出，各項模式符合度指標皆不理想，例如：因素負荷量沒有介於 0.5~0.95 之間，卡方值與自由度比例大於 3，GFI、AGFI、NFI、IFI 等整體符合度指標皆小於 0.90，以及各變項的信度皆不理想。顯示初始模式有許多不太恰當之模式假設，應予以修改。本研究根據上述分析，將創造歷程之變項修改後形成模式二，再將創造成品之變項修改後形成模式三。最後再根據修改指標之建議與研究團隊討論之結果修改結構模式（潛在自變項與潛在依變項間的關係），形成最終模式。

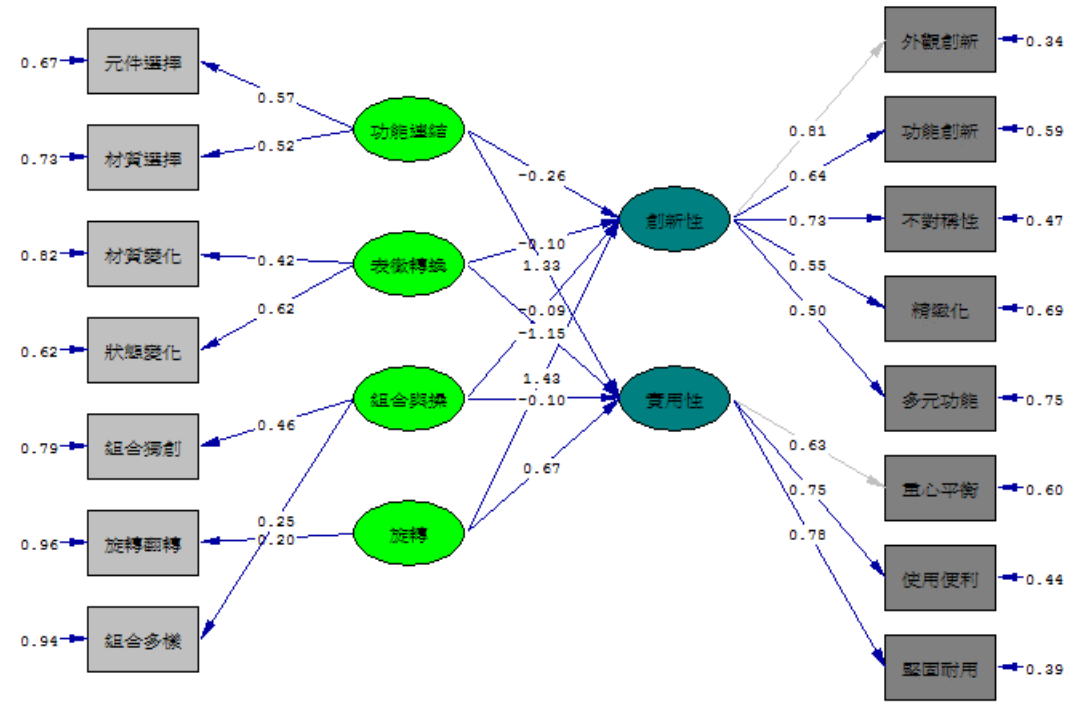

圖 7 創造歷程與創造成品的初始模式分析結果

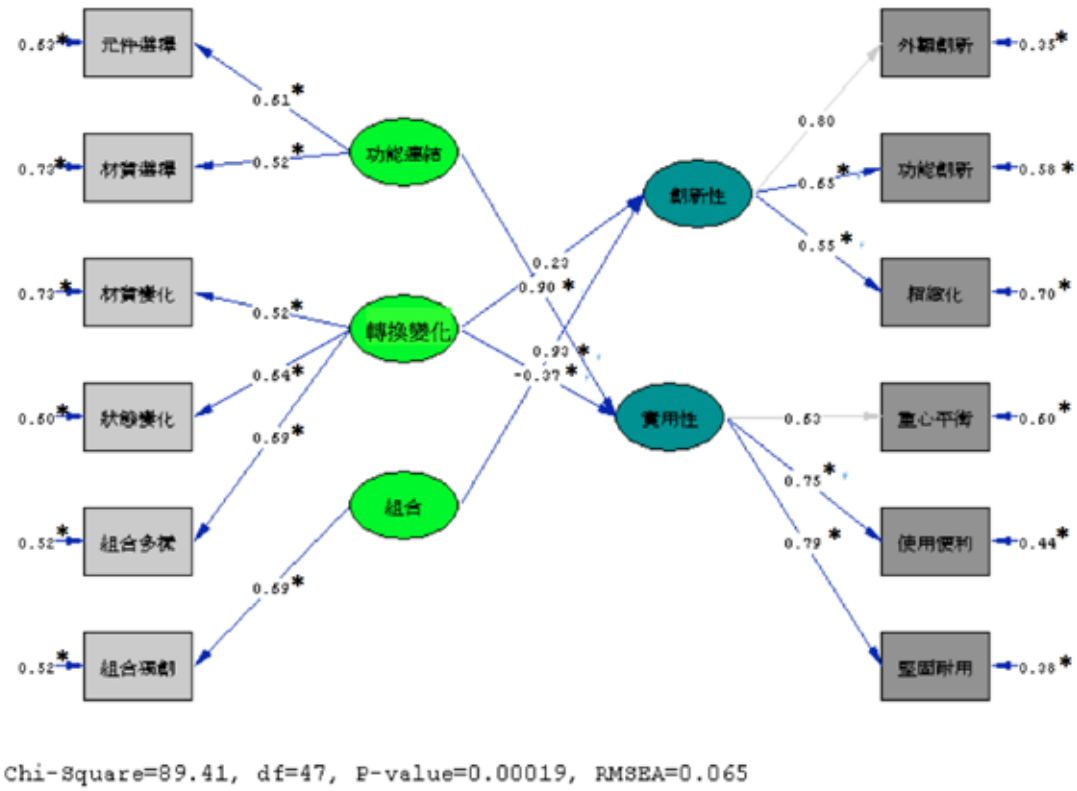

圖 8 創造歷程與創造成品的最終模式分析結果

該測驗之最終模式參數分析結果如圖 8 所示。從歷程評分與成品評分的測量模式中可以看出，所有測量模式的因素負荷量皆達顯著，各項評分指標之 λ 值皆大於 0.5，且 t 檢定皆顯著不等於零（p<0.05）。但仍有將近一半的測量指標其測量誤差接近 0.5。由於本研究乃以人為評分來建立指標分數，因此其測量信度不如客觀答題反應的量表，未來需要加強評分者的訓練以提高測量信度。

從結構關係模式來看，創造歷程之「功能連結」對創造成品「實用性」的迴歸係數達 0.90，顯示此歷程可以提升創造成品之「實用性」；而創造歷程之「組合方式」對創造成品「創新性」的迴歸係數達 0.93，顯示此歷程可以提升創造成品之「創新性」。創造歷程之「轉換變化」對創造成品「創新性」有正向影響（0.23）、但對創造成品「實用性」則有負向影響（-0.37），顯示「轉換變化」會提升創造成品之「創新性」，但會降低創造成品之「實用性」。不過此歷程對創造成品「創新性」的影響係數未達顯著，未來還需要更多創造歷程的評量指標來了解其對創造成品創新性的關係。

最終模式之模式符合度資料如表五所示。從表 5 中可以看出，各項模式符合度指標大致能符合要求，例如：測量模式之因素負荷量都介於 0.5~0.95 之間，卡方值與自由度比例小於 3，GFI、AGFI、NFI、IFI 等整體符合度指標皆大於 0.90 等。顯示最終模式已經與實際資料大致相符。雖然「改變轉換」對創造成品「創新性」的影響未達顯著，但根據過去相關研究（李珮瑜，2011），創新性高低不同者，其在「變化轉換」的相關指標上皆有顯著差異，因此該項迴歸線予以保留。

表 5　初始模式與最終模式之模式符合度指標

	模式符合度指標	初始模式	最終模式
基本適配度指標	沒有負的誤差變異	是	是
	參數間的相關絕對值不會太接近 1	不符合	是
	因素負荷量 0.5-0.95	不符合	是
整體模式適配度指標	χ 2 未達顯著	251.92 （P=0.0000）	96.26 （P=0.0002）
	χ 2 比率 <3	251.92/76=3.31	是（96.26/47=2.04）
	GFI 指數 >0.9	0.86	是（0.94）
	AGFI>0.9	0.77	0.89
	RMSEA<0.05	0.11	0.065
	RMR/SRMR<0.05	0.092	0.071
	Q-plot 殘差分佈斜度 >45	否	是
	NFI>0.9	0.83	是（0.90）
	IFI>0.9	0.87	是（0.94）
	NHFI>0.9	0.82	是（0.92）
模式內在品質	個別項目的信度 >0.5	否	半數符合
	潛在變項平均變異量 >0.5	否	否
	估計的參數都達顯著	否	大致符合（1 項不符）
	修正指標是否小於 3.84	否	是

四、結論

本文主要介紹當代電腦化測驗之題型，並參考 Parshall 與 Harmes（2007）的分類向度舉例說明近代幾項創新的電腦化測驗題型。以目前資訊科技之進度速度來看，電腦化測驗題型種類繁多，也頗接近日常生活的情境，讓受測者可以使用許多便利的作答方式來展現其潛在特質，提高了測量工具的效度。

然而，電腦化測驗題型僅是蒐集受測者外顯行為之工具，測驗的主要目的仍然是評估受測者的潛在特質，而非只是蒐集外顯行為。因此，測驗編製者仍須根據測驗的理論架構與所使用的電腦化測驗題型的特性，尋找適當的測量模式，並驗證該測量模式是否能有效地解釋受測者的作答反應，此即建構效度的驗證過程。若測量模式所預期的結果與受測者實際作答反應的吻合度不佳，則需考慮調整題型或調整測量模式。當測量模式能有效解釋受測者的作答反應時，創新的電腦化測驗題型才能真正發揮測量的功能。

致謝

本文感謝教育部「邁向頂尖大學計畫」與科技部「跨國頂尖研究中心計畫」（NSC103-2911-I-003-301）支持。

參考文獻

林正堅（2011）。汽車駕訓之動感模擬系統設計。**勤益新聞網**。2013 年 5 月 31 日檢索。網　址：http://enews.ncut.edu.tw/web/news/news.php? NewsCategoryID=13 & NewsID=127&Year=2011

李佩瑜（2011）。**潛在類別分析與二階段群集分析分群效果之比較研究**。國立臺灣師範大學教育心理與輔導學系碩士論文。

葉玉珠（2004）。科技創造力測驗的發展與常模的建立。**測驗學刊，51**（2），127-162。

張國恩、劉遠禎、王曉璿、蕭顯勝、宋曜廷、陳柏熹（2008）。**中小學資訊能力評量機制發展與推廣計畫結案報告**。教育部。

陳柏熹（2008）。**多元智慧電腦化適性測驗的發展：不同 MIRT 模式之適配度比較**。**第 47 屆**台灣心理學會，台北。十月。

陳柏熹（2011）。**創造力電腦化測驗系統之發展暨學生創造歷程之研究：IRT 取向（第二年）**。國科會期中報告（NSC 98-2511-S-003-011-MY3）。

陳柏熹、歐詠芝（2011）。肢體動作電腦化測驗的研發與效度研究。**2011 心理與教育測驗學術研討會**。台北。十月。

陳柏熹（2014）。電腦化大學生基本素養測驗簡介。**評鑑雙月刊，第 47 期**。

蕭顯勝、林建佑、邱敬尊（2009）。應用擴充實境技術發展電腦化動作技能測驗系統。第五屆台灣數位學習發展研討會論文（TWELF 2009）。台灣：台南大學，9 月。

Amabile, T. M. (1996). *Creativity in context.* CO: Westview.

Andrich, D. (1978). Rating formulation for ordered response categories. *Psychometrika*, 43, 561–573.

Birnbaum, A. (1968). Some latent trait models and their use in inferring an examinee's ability. In F. M. Lord & M. R. Novick, *Statistical theories of mental test scores.* Reading, MA: Addison-Wesley.

Chen, K.-L., Huang, S.-H., Liu, S.-Y., Chen, P.-H. (2014). Energy Literacy of Secondary Students in Taiwan: A Computer-based Assessment. The Third International Conference on E-Learning and E-Technologies in Education (ICEEE 2014), Malaysia: Kuala Lumpur.

Chen, P.-H., Hsu, C.-Y., Huang H.-Y., Li, P.-W., Chen Y.-S., Yang C.-Y., Yeh, T.-T. (2013). Development of the general literacy test for university students. Paper presented at the 78th International Meeting of Psychometric Society 2013 (IMPS 2013). Amsterdam, Netherland.

Chen, P. H., Kuo, J. W., & Sung, Y. T. (2011). Influence of Pre-Test Design on the Precision of the Parameters Estimation in the Multidimensional Items Bank. Paper presented at the 76th Annual Meeting of the Psychometric Society (IMPS 2011). Hong Kong, China.

Fischer, G. H. (1973). The Linear logistic model as an instrument to educational research. *Acta Psychologica*, 37, 359-374.

Hattie, J. (1981). *Decision criteria for determining unidimensional and multidimensional normal ogive models of latent trait theory.* Armidale, Australia: The University of New England, Center for Behavioral Studies.

Jöreskog, K. G., & Sörbom, D. (1993). *LISREL 8: Structural equation modeling with the SIMPLIS command language.* Mooresville, IN: Scientific Software, INC.

Linacre, J. M. and Wright, B. D. (1993). *A user's guide to FACETS: Rasch-measurement computer program*. Chicago, IL: MESA Press .

Liu, T. C., Kinshuk, Lin, Y. C., & Wang, S. C.(2012). Can verbalisers learn as well as visualisers in simulation-based CAL with predominantly visual representations? Preliminary evidence from a pilot study. *British Journal of Educational Technology,* 43(6), 965-980.

Lord, F. M. (1952). A theory of test scores. Psychometric Monograph, No. 7.

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

Mckinley, R. L., & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods & Instrumentation,* 15, 389-390.

Parshall. C. G. & Harmes, J. C. (2007). Designing templates based on a taxonomy of innovative items. (2007). In D. J. Weiss (Ed.). *Proceedings of the 2007 GMAC Conference on Computerized Adaptive Testing.* Retrieved 2014/5/25 from www.psych.umn.edu/psylabs/CATCentral/

Parshall. C. G., Davey, T. & Pashley. P. J.(2000). Innovative Item Types for Computerized Testing. In Wim J. van der Linden and Gees A.W. Glas (Eds.). *Computerized Adaptive Testing: Theory and Practice,* pp 129-148.

USMLE (United States Medical Licensing Examination).(2014). Tutorial and Practice Items for Multiple Choice Questions and Primum Computer-based Case Simulations (CCS). Retrieved 2014 from http://www.usmle.org/practice-materials.

Wainer, H., Bradlow, E. T., & Du, Z. (2000). Testlet response theory: An analog for 3PL model using in testlet-based adaptive testing. In W. van der Linden & C. A. W. Glas (Eds.), *Computerized Adaptive Testing: Theory and Practice.* (pp. 245-269). London: Kluwer.

Wang, W.-C. & Wilson, M. R. (2005). The Rasch testlet model. *Applied Psychological Measurement.*

Wang, W.-C., Wilson, M. R., & Adams, R. J. (1997). Rasch models for multidimensionality between items and within items. In M. Wilson, G. Engelhard & K. Draney (Eds.), Objective measurement: *Theory into practice.* (Volume 4, pp. 139-155). Norwood, NJ: Ablex.