第五章　雛形系統成效初步評估

本研究的系統成效評估，則包括「線上測驗與智慧評分子系統」及「輔助教師產製測驗題庫子系統」等二個部分之成效評估。「線上測驗與智慧評分子系統」之成效評估除了採用與其他現有系統直接進行功能比較外，另外也將採用實驗法來進行測驗評分機制的評分效力檢驗，以期望能瞭解本研究提出的智慧評分機制概念是否具備與傳統教師紙筆測驗評分相同的評分效力。而「輔助教師產製測驗題庫子系統」之成效評估，除了針對教師人工出題與電腦產製試題的方式進行關鍵績效指標的成效比較外，另外在電腦產製的試題品質評估方面，則請教師實際操作系統，再以問卷及訪談方式探討系統的成效。

第一節 輔助教師產製測驗題庫子系統之系統成效評估

由於目前現有線上測驗系統的測驗題庫試題內容大多均由人工進行產製，因此教師往往需要耗費大量人力與時間投入至試題的產製工作，而且教師在產製試題時，通常也未考量每個試題所能評量的知識類型與認知層次，因此本研究將以電腦輔助教師產製的試題與教師人工產製試題的方式，進行關鍵績效指標的評估。所以在系統的關鍵績效指標部份，將依據本研究之目的進行選用。經由先導實驗中的研究結論顯示，在 E-Learning 環境中若要以人工方式進行大量試題題庫的產製，則是一件非常困難的工作，特別是在題目內容變化與數量上容易受限。以教師人工方式產製電腦可自動評分的是非題、單選題、複選題及填充題時，試題能用以評量的知識類型涵蓋 Bloom 分類中的事實知識、概念知識及程序知識等三種層次，而試題能用以評量的認知能力則涵蓋 Bloom 分類中的記憶、了解、應用、分析及評鑑層次。因此，本研究在進行系統輔助教師產製試題的成效評估時，選用的系統關鍵績效指標將包括試題產製效率、試題變化、試題可涵蓋 Bloom 分類的知識層次及認知層次範圍、符合試題出題原則、試題品質、防止學生記憶答案能力、防止同一試卷相似題目重複出現能力等。

本研究對於「輔助教師產製測驗題庫子系統」的系統成效評估以「系統輔助教師產製試題成效之實證研究」及「與教師人工產製試題及現有測驗系統進行功能績效比較」等二個部分進行。

一、 系統輔助教師產製試題成效實證研究之研究設計

(一) 問題陳述：本研究將透過國內資訊管理系教師親自操作本研究的系統，並進而針對本研究之系統其所能達到的出題效益，進行成效評估。

(二) 研究樣本：本研究以國內 11 所大專院校共 14 位資訊管理系教師來進行實驗，這些教師均是已全程參與本研究先導實驗的教師成員，在先導實驗中原本有 15 位教師參與，但由於先導實驗共花費二個月時間，且需要耗費大量時間與心力參與，因此其中有 1 位教師在完成先導實驗後因時間無法負荷，導致無法參與最後的系統成效評估實驗。所有參與本實驗的教師均曾經任教過管理資訊系統或資訊管理導論課程，並將協助瞭解「輔助教師產製測驗題庫子系統」的系統成效。

(三) 資料收集方法：採 Likert 5 點區間尺度量表進行問卷調查及訪談，以取得研究過程中所需的資料。

(四) 實驗流程：本實驗在開始進行前，先錄製了試題產製過程的展示說明影片(含系統

產製試題原理)，然後讓參與實驗的教師實際操作系統的主要功能，並檢視系統輔助教師產製試題的結果、實際操作系統自動輔助試卷選題功能，隨後電腦從所有電腦產製的試題中，隨機產生知識概念不重複的試題共 50 題(是非 10 題、單選 20 題、複選 10 題、填充 10 題)，最後再由參與實驗的教師填寫 Likert 5 點區間尺度的電腦產製試題合適性及採用意願問卷量表(量表內容請參考附錄四)，以及電腦產製試題子系統的系統成效問卷量表(量表內容請參考附錄五)。當問卷資料回收後，則針對每位教師所填寫的問卷內容，及與其他教師有不同看法的部分進行深入訪談，以瞭解每個教師對於本研究系統成效的真實看法。

二、 系統輔助教師產製試題成效實證研究之資料分析

(一) 試題類型分佈分析

本研究經過「輔助教師產製測驗題庫子系統」的建置後，也讓電腦針對先導研究過程中，參與實驗教師所使用的第二章教材內容進行試題產製，最後經由電腦所產製試題之題型分佈結果如表 21 所示，產製完成的試題中包括了是非題 1153 題，單選題 6612 題，複選題 10659 題，填充題 197 題，合計 18621 題，這些題目共來自於 279 種知識概念；而在先前的先導實驗中，15 位教師共同產製的試題是非題 143 題，單選題 106 題，複選題 68 則包括了題，填充題 69 題，合計 386 題，顯然本研究在各題型的試題數量，均比教師人工出題的數量更多。

若從 Bloom 認知目標向度來分析，電腦產製試題之 Bloom 分類向度分佈結果如表 22 所示。就知識向度而言，電腦輔助教師產製的試題所評量之知識類型包括了事實知識 12507 題，概念知識 3002 題，程序知識 3112 題；而在先前的先導實驗中，15 位教師共同產製的試題所評量之知識類型包括了事實知識 276 題，概念知識 98 題，程序知識 12 題。就認知歷程向度而言，電腦輔助教師產製的試題所評量之認知類型包括了記憶層次 3058 題，了解層次 32 題，應用層次 50 題，分析層次 14968 題，評鑑層次 513 題；而在先前的先導實驗中，15 位教師共同產製的試題所評量之認知層次包括了記憶層次 260 題，了解層次 52 題，分析層次 73 題，評鑑層次 3 題。電腦輔助教師產製試題所涵蓋的 Bloom 分類範圍，與先前教師人工出題的範圍相同，在是非題、單選題、複選題及填充題等四種題型的限制下，認知歷程中的創造層次試題均難以發展。而就認知歷程而言，無論是電腦自動出題或是教師人工出題，「記憶」層次通常以「事實知識」為主，「了解」層次通常以「概念知識」為主，而「應用」層次通常以「程序知識」為主，此結果與陳筱菁(民 93)的研究發現相同。本研究中因為試題產製規則中能進行大量的對照詞彙型試題及相關整併型試題，此二類試題因為能產生大量的單選題及複選題，也能產生認知歷程中的分析層次辨別能力試題，因此本研究中電腦輔助教師產製的試題，在單選題、複選題及分析層次上的數量比例會有較高的情形。

表 21: 電腦輔助教師產製試題之題型分佈結果

試題題型	是非題	單選題	複選題	填充題	小計題
刪除重複試題	6.19% **(1153)**	35.51% **(6612)**	57.24% **(10659)**	1.06% **(197)**	100% **(18621)**

表 22: 電腦輔助教師產製試題之 Bloom 分類向度分佈結果

知識向度	認知歷程向度					
	記憶	了解	應用	分析	評鑑	小計
事實知識	2086(11.20%)	0 (0%)	0 (0%)	10421(55.96%)	0 (0%)	12507(67.17%)
概念知識	944(5.07%)	32 (0.17%)	0 (0%)	2026(10.88%)	0 (0%)	3002(16.12%)
程序知識	28(0.15%)	0 (0%)	50(0.27%)	2521(13.54%)	513(2.75%)	3112(16.71%)
小計	3058(16.42%)	32 (0.17%)	50(0.27%)	14968(80.38%)	513(2.75%)	18621(100%)

當系統產製完成試題產製工作後，線上測驗系統內便存在許多的測驗題庫，教師可以透過系統中的自動選題編製試卷功能，來設定教師所欲產生的試卷組合，教師可以選擇以 Bloom 知識層次、Bloom 認知層次、測驗題型為評量概念基礎的試題，例如圖 66 為某一份試卷的試題，從知識類型向度而言，此試題包括 8 題事實知識、5 題概念知識、2 題程序知識的試題；從認知歷程向度而言，此試題包括 8 題記憶層次、3 題了解層次、3 題分析層次及 1 題評鑑層次的試題。當學生完成測驗後，系統除了產生試題內容的評分及知識回饋結果外，也會提供如圖 66 所示的學習成效能力指標圖。圖 66 顯示出某位學生的試題測驗中，8 題事實知識答對了 7 題，程序知識及概念知識答錯的比例也高達 40%以上，而在 3 題知識瞭解能力的試題上完全都答對，在評量較高層次的分析能力及評鑑能力則成果不佳，因此學習者可藉由學習成效能力指標圖，來瞭解自己在各種不同知識層次或認知層次能力上的表現情形，並針對表現不好的類型試題，進行學習方法的修正。

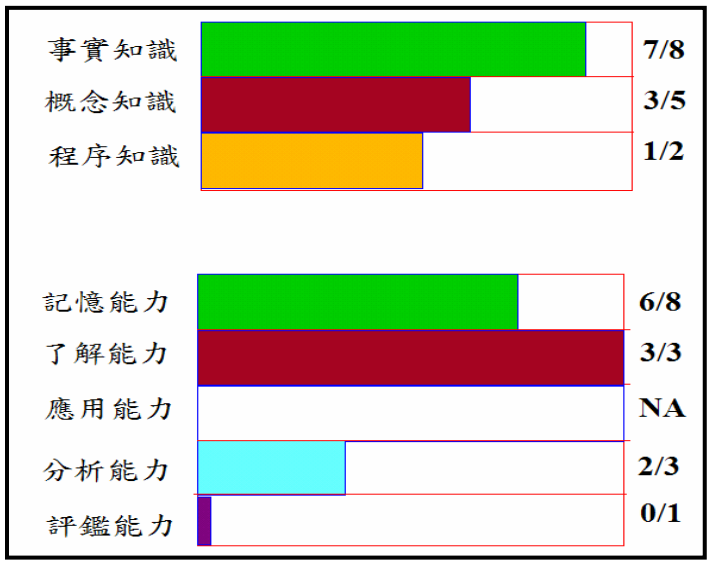

圖 66: 學習成效能力指標圖

(二) 電腦輔助教師產製試題之試題採用意願分析

本研究從電腦輔助教師產製完成的 18621 題試題中(共有 279 個知識概念)，進行二階段隨機抽樣，以產生知識概念不重複的試題共 50 題(是非 10 題、單選 20 題、複選 10 題、填充 10 題)。由於在本實驗研究中，電腦輔助教師產製完成的 18621 題試題，主要來自於 279 個不同的知識概念，因此每一個原始教材中的知識概念，可能產生許多不同的試題內容變化及不同的題型試題。所以在第一階段的隨機抽樣中，主要是從 279 個知識概念中，隨機抽取出 50 個知識概念；而第二階段的隨機抽樣中，則是針對第一階段抽出的

每個知識概念，將其所屬的不同試題內容及題型試題中，再隨機抽取出其中一個試題。最後再由參與實驗的 14 位資管系教師針對每一個試題的可用性、合理性、評量性及願意採用該試題的認同度，填寫 Likert 5 點區間尺度的試題品質量表，若參與教師給予該試題 1 分的認同度，表示教師認為該試題品質不佳，因此不認同此試題可以被使用在正式的測驗中；教師若給予 5 分，則表示教師認同該試題的品質，而且非常願意採用該試題於正式的測驗中。

本量表之 Cronbach α 值為 0.975，量表中的每個問項題目皆透過教材知識及試題產製規則產生，並經過二位領域教師檢視，因此具有高度的信度與內容效度。

所有參與教師填答的試題採用意願問卷結果彙整如表 23~26，表 23 顯示參與實驗的教師對於由電腦隨機選出的 10 題是非題，其採用試題的意願分數平均為 4.06 分；表 24 顯示參與實驗的教師對於由電腦隨機選出的 20 題單選題，其採用試題的意願分數平均為 4.12 分；表 25 顯示參與實驗的教師對於由電腦隨機選出的 10 題複選題，其採用試題的意願分數平均為 4.14 分;表 26 顯示參與實驗的教師對於由電腦隨機選出的 10 題填充題，其採用試題的意願分數平均為 4.14 分。由表 23~26 的結果顯示，大多數教師對於電腦產製是非題、單選題、複選題及填充題的試題內容，普遍均認同這些試題的可用性、正確性、合理性及評量性，因此願意採用這些經由電腦輔助教師產製的試題。

雖然從表 23~26 可以看到大多數教師對於電腦輔助教師產製各類試題的試題內容，給予正面的認同，但是在表 23~26 中也可以看到標準差的值均在 1 左右，而且許多試題也有不同教師分別給予 1 分及 5 分的分數，這顯示出不同的教師對於相同的題目內容，可能有不同的認同度。因此本研究針對每個試題有填寫不認同的教師進行訪談，以期能挖掘出實際的試題問題。

研究者發現實驗代號 E20 的教師，對於所有是非題給予 3 分之認同評價，而在單選題、複選題及填充題部分則給予 4~5 的高認同評價，在與 E20 教師訪談後發現，該教師因個人教學經驗及個人長期對於測驗題型的看法感受，相當排斥在測驗中使用是非題，該教師認為許多的知識議題並不適合使用單純的是非題來評量，而且容易造成學生對於知識的理解變成只是對與錯的認知，而無法深入瞭解知識的意涵。E20 教師也指出，他在給予是非題的試題認同評價時，並非真的覺得試題內容不恰當，而單純是認為測驗中使用是非題是不恰當的，因此由個人對於測驗題型的喜好感受而影響對於是非題型的認同評價。

另外本研究也發現了另一個相當對比的現象，代號 D14 的教師則對於所有填充題給予 1-2 分之低認同評價，而在是非題、單選題及複選題部分，則給予 3~5 分的評價，在與 D14 教師訪談後發現，該教師反而相當排斥填充題型，並認為填充題會限縮了學生的思考空間，因此該位教師相當排斥測驗中使用填充題試題，而事實上該名教師也表示，他對於電腦輔助教師產製的填充題試題內容並沒有意見，但因為本研究直接詢問該教師對於每個填充試題是否適合被用以進行測驗，且參與實驗教師是否願意採用該試題時，因為 D14 教師根本就非常排斥使用填充題型的試題，所以如同 E20 教師一樣，教師個人對於測驗題型的喜好感受會直接影響對於不同題型的認同評價。此外，D14 教師相當堅持試題中應該採用問答題等試題，因此許多課程可能不適合透過 e-learning 環境來進行學習成效的評量，該教師並且認為單純考記憶層次的回憶認知能力的試題可能沒有意義。但是事實上，在教育理論及 Bloom 認知目標理論中描述，記憶層次可區分為記憶確認及

記憶回憶，其中的記憶確認是指從現有知識中確認出與先前所學習知識一致的內容，而記憶回憶則是能將先前所學習的重複的唸出或寫出的能力。每一種認知層次的試題，都是用來評量不同的知識認知能力，此外，若要讓學生有足夠能力回答問答題的試題，對於一些課程領域中的重要專有名詞或基本元素概念，學生仍有必要將這些知識記憶下來，因此之前也有教授認為目前有些學界教師過份重視問答題的功效，卻忽視了許多學生對於基本知識及名詞瞭解的缺乏事實，若直接期望學生進行更高層次的知識評論，似乎沒有考量到學生是否已經有其他認知歷程的能力程度，這值得教育學者進一步的重視與探討。

而代號 A11 的教師對於填充題有比較多的疑慮，該教師認為透過線上測驗進行填充題的測驗，每個填充題答案可能有許多可替換的詞彙，因此容易造成學生的誤解，而電腦是否能有效的處理填充題的評分則是一個很大的問題。但由於本研究再讓老師進行電腦產製試題的可用性及認同評價時，僅希望教師針對試題本身進行評價，因此代號 A11 已將其個人對於電腦處理填充題評分的能力認知，一併做為評價試題的準則。但在研究者告知本研究的系統已經能夠處理填充題的語意模糊評分能力時，則該教師對於電腦產製試題的品質則有更高的評價，但表 23~26 的評價彙整仍為每位老師最原始的評價值，而好幾位教師在經過訪談後，均希望能調高對某些試題的評價，並消除個人對於試題題型主觀認知的部分，然而本研究則僅希望將老師們的想法能被忠實的呈現，因此未讓參與實驗的教師進行原始評價分數的修改。另外，A11 教師也認為選擇題的備選答案雖然可以有多樣化的選擇，但是不同教科書的版本可能會有所出入，因此出題時並不一定要侷限在標準特定答案內，這也是教師人工出題或是電腦輔助教師產製試題都要考慮的問題。

而代號 D26 的教師對於記憶類型試題均給予較低的認同評價，且對於術語的記憶性題目，D26 教師較偏好以單選題方式出題，而不喜歡以是非題方式出題，因此該教師也認為其實對於各試題的認同評價，其實會有個人的主觀性與偏見，不同的教師應該會有不同的看法與喜好。

而代號 A11、A26 的教師則認為教師對題目的認同觀點，會受到自己對於課程目標的定義及個人對於題型的主觀偏好而有差異，就如同全國性的統一入學測驗，也是會有許多人對於不同的題目給予不同的看法與評價，因此對於試題的評價僅能表現出教師個人對於試題的偏好，因此既使給予某些試題較差的認同評價，有時候並非是試題本身的品質有顯著的問題。

代號 A31 教師則認為電腦產製的試題，已經能包含了傳統教師所能發展的一般性基本試題，所涵蓋的試題範圍也很廣，但是若是要發展概念知識性的題目，則由教師親自設計試題，其試題品質會更好。此外，電腦產製的複選題內容，其中有些題目要求選出不正確答案的複選題選項與課文中的答案很接近，不容易判斷選出答案，然而備選答案是否要選擇容易混淆但相近的備選答案來考倒學生，還是要讓學生可以很容易選擇出來，這牽涉到題目的難易度與鑑別力，A31 教師則偏向於能夠不要考倒學生，也不需要這麼重視題目的鑑別能力，而只需要確認學生是否對於該瞭解的知識都能清楚瞭解即可。另外，A31 教師也指出藉由電腦科技的輔助，利用電腦可輔助教師產製大量、豐富多樣的試題，也可以補強人工出題費時費力的缺點，同時又能提高試題的可信度與效能。

綜合上述老師的看法，本研究認為不同的教師對於特定題型及題目難易度會有特殊

的認知偏好，因此會造成某些教師對於特定題型給予試題較差的認同分數，但整體而言，參與實驗的教師對於電腦產製的試題合適性與認同度方面，均給予正面的評價，顯示本研究的試題能被一般教師所接受及使用。

然而，在本研究中光是針對特定章節教材，電腦便已輔助教師產製 18621 個試題，並具有輔助教師快速建立大量題庫的能力。受限於研究時間及研究目的，本研究僅從 18621 個試題中隨機抽取出 50 個試題，讓多位教師來評估電腦產製試題的可用性，因為採用二階段隨機抽樣方式抽取試題，因此本研究雖然僅抽取出 50 個試題讓教師評估，但因符合統計抽樣的理論觀念，所抽取的試題樣本也能具備題庫母體的特徵及代表性。

表 23: 參與教師對於電腦產製試題的是非題採用意願結果

	是1	是2	是3	是4	是5	是6	是7	是8	是9	是10
試題平均	3.64	4.00	4.36	4.50	4.36	4.36	4.21	3.71	3.79	3.71
試題S.D.	0.93	0.96	0.74	0.76	0.74	0.84	0.97	1.20	1.05	1.20
MIN	2	3	3	3	3	3	2	2	2	2
MAX	5	5	5	5	5	5	5	5	5	5
平均	4.06									
標準差	0.98									

表 24: 參與教師對於電腦產製試題的單選題採用意願結果

	單1	單2	單3	單4	單5	單6	單7	單8	單9	單10	單11	單12	單13	單14	單15	單16	單17	單18	單19	單20
試題平均	4.5	3.8	4.3	3.7	4	4.3	4.4	4	4.3	4.2	4	4	4.6	3.9	4.4	3.9	4.4	4	4.3	3.6
試題S.D.	0.8	1.4	0.6	1.3	1.4	0.8	0.7	1	0.8	0.7	0.9	0.9	0.7	0.9	0.9	1.1	0.8	1	0.9	1.2
MIN	3	1	3	1	1	3	3	2	3	3	2	2	3	2	2	1	3	2	3	2
MAX	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5	5
平均	4.12																			
標準差	0.91																			

表 25: 參與教師對於電腦產製試題的複選題採用意願結果

	複1	複2	複3	複4	複5	複6	複7	複8	複9	複10
試題平均	3.64	4.21	4.21	3.93	4.14	4.57	4.29	3.86	4.21	4.29
試題S.D.	1.15	0.97	0.89	0.83	1.10	0.65	0.83	1.17	0.97	1.14
MIN	2	2	2	3	2	3	3	2	2	2
MAX	5	5	5	5	5	5	5	5	5	5
平均	4.14									
標準差	0.98									

表 26: 參與教師對於電腦產製試題的填充題採用意願結果

	填1	填2	填3	填4	填5	填6	填7	填8	填9	填10
試題平均	4.14	4.21	4.07	4.36	4.21	4.36	4.14	3.93	3.93	4.07
試題S.D.	0.95	0.89	1.14	1.01	0.97	1.01	1.17	1.33	1.44	1.33
MIN	2	2	2	2	2	2	2	1	1	1
MAX	5	5	5	5	5	5	5	5	5	5
平均	4.14									
標準差	1.11									

表 27: 參與教師對於電腦產製試題子系統具備的系統能力評估結果

電腦產製試題子系統具備的系統能力評估項目	Mean	S.D
01.產製評量「事實知識」的試題能力	4.43	0.65
02.產製評量「概念知識」的試題能力	3.93	0.73
03.產製評量「程序知識」的試題能力	4.07	0.73
04.產製評量各知識類型的試題能力	4.21	0.80
05.產製評量知識「記憶」的試題能力	4.29	0.83
06.產製評量知識「了解」的試題能力	4.07	0.92
07.產製評量知識「應用」的試題能力	4.00	0.88
08.產製評量知識「分析」的試題能力	3.64	1.15
09.產製評量知識「評鑑」的試題能力	3.79	0.89
10.產製評量知識各類認知的試題能力	3.93	0.73
11.產製「是非題」的試題能力	**4.50**	0.65
12.產製「單選題」的試題能力	4.43	0.65
13.產製「複選題」的試題能力	4.00	0.96
14.產製「填充題」的試題能力	4.21	0.98
15.試題具備產生 Bloom 概念資訊的能力	4.14	0.86
16.建立龐大題庫，系統比人工出題更有效率	**4.71**	0.47
17.建立龐大題庫，系統產製的試題數量能力較人工出題佳	**4.57**	0.65
18.產製的試題變化，比人工出題更有能力	4.21	0.89
19.試題數量足以提供給線上學習系統的測驗評量使用	**4.50**	0.65
20.試題包含教材中所有重要的知識概念	4.43	0.65
21.試題與老師產製的試題重點類似	4.14	0.53
22.針對相同知識概念，產生不同試題變化組合的能力	**4.50**	0.76
23.**試卷選題功能**使試卷內不會有評量相同知識概念的重複試題	4.36	0.50
24.能達成是非、單選、複選及填充題出題工作	4.21	0.89
25.能協助教師發展一般常見的基本層次測驗試題	**4.50**	0.52
26.我會想要利用此系統來協助產製大量題庫及選取試卷題目	4.36	0.84
27.試題符合試題編製理論的原則要求	4.21	0.80
28.試題包含教師沒有想到的試題內容	4.21	0.80
29.試題品質與教師第一階段編製的試題品質類似	4.00	0.68
30.試題都不同，且都具有特定的意義及代表性	4.03	0.92

(三) 電腦產製試題子系統的系統成效評估分析

此部分之實證研究主要的目的在於瞭解電腦輔助教師產製試題的系統成效，因此本研究請參與實驗的教師在實際操作系統後，填寫「電腦輔助教師產製試題子系統的系統成效問卷量表」，此量表問項參考測驗系統及編製試題原則等相關文獻之指標，並經二位領域專家針對問卷內容及詞句給予意見，因此具備相當高的內容效度。另外在信度部分，本量表的 Cronbach α 值為 0.973，因此量表具備高信度。

電腦產製試題系統具備的系統能力評估項目結果如表 27 所示，參與教師普遍認為利用電腦能夠產製出用以評量「事實知識」、「概念知識」及「程序知識」等各種知識類型的試題，但對於電腦產製用以評量「概念知識」的成效上不如「事實知識」的效果好，此部分有待後續研究針對「概念知識」類型的試題產製，進行產製規則的研究。而利用電腦也能夠輔助教師產製出用以評量認知歷程的「記憶」、「了解」、「應用」、「分析」及「評鑑」等類型的試題，其中透過電腦輔助教師產製用以評量較低層次認知的試題表現較佳，對於更高層次的「分析」及「評鑑」試題，則表現的較不理想。這個結果也顯示，評量越高層次的認知類型試題，較不容易透過基本的結構來讓電腦產製，因此更高層次的試題仍須由教師人工來輔助為宜。

在產製試題的題型能力方面，教師普遍認為無論在是非題、單選題、複選題及填充題，電腦均能進行有效的試題產製，特別是是非題的試題產製表現最佳，複選題的試題產製為此四種類型中表現較差者，但整體而言均得到參與實驗教師的肯定。

此外，研究結果也顯示出透過電腦產製的試題，所具備的能力尚包括：具備產生 Bloom 概念資訊的能力、試題變化較人工出題佳、試題涵蓋教材中所有重要知識概念、試題與教師人工產製的重點類似、試卷選題功能使試卷內不會有評量相同知識概念的重複試題、能達成電腦可自動評分題型的出題工作、試題符合試題編製理論的原則要求、試題包含教師沒有想到的試題內容、每個試題都具有特定的意義及代表性、試題品質與教師第一階段編製的試題品質類似。而在建立龐大題庫能力方面，先前在前導研究中 15 位教師平均花費 3.9 小時，共產製了 386 題試題，電腦僅在 5 分鐘內便自動完成 18621 題的試題。因此大多數老師均認為系統比人工出題更有效率，而且試題的數量也較人工出題佳，所以電腦產製的試題足以提供給線上學習系統使用。另外，老師們也認為電腦輔助教師產製試題能協助教師發展一般常見的基本層次測驗試題，因此老師們會想利用此系統來協助其產製大量題庫及選取試卷題目，而教師所節省的時間，則可進行更高層次的試題發展。

在本研究發現無論是教師人工出題或電腦輔助教師產製試題，均無法針對「創造」認知層次的試題進行是非題、單選題、複選題及填充題的發展，因此有必要針對管理資訊系統領域的「創造」層次試題進行探討，以確認「創造」層次試題的確會因為電腦測驗環境所支援的測驗題型而受到限制。依據 Bloom 分類理論對於「創造」認知層次的定義，是指學生能將知識要素聚集在一起形成一個協調或具有整體功能的能力，此類別項目又再分為「產生」、「規劃」及「製作」等三種子類別能力。「產生」能力是指學生能呈現問題並達成符合某些準則的多種選擇方向或假設能力，在管理資訊系統領域中，此類型的試題範例可以是「請針對 XX 公司個案中對於資訊系統取得方法的決策，是否正確？如果不正確，請提出你認為不正確的原因，並說明您的看法與理由」。「規劃」能力是指學生能規劃一個符合問題準則的解決方法能力，在管理資訊系統領域中，此類型的試題範例可以是「請依照下方所列之需求，規劃一個具備醫療診斷專家系統能力的系統架構，並說明架構中每個元素的功能與運作方式。」「製作」能力是指能執行而得以解決符合某些特徵要求的問題能力，在管理資訊系統領域中，此類型的試題範例可以是「請依照進位轉換公式的概念，實際撰寫一個程式，來達到 2 進位、8 進位、10 進位及 16 進位資料相互轉換的功能。」由上述針對管理資訊系統領域的「創造」認知能力的試題範例中可以得知，「創造」層次著重於學習者個人在知識內化後的整合概念創造能力，因此無法透

過是非題、單選題、複選題等認識型試題來評量，也難以透過填充題等建構型試題來讓學習者進行知識概念整合後的觀點論述。因此，既使在其他領域中，有關於「創造」層次能力的評量，仍適合利用問答申論或實作題等方式來進行，但這類型試題目前仍無法透過資訊技術來自動評分，所以這也是本研究中無論教師人工出題或電腦輔助教師產製試題均無法透過是非題、單選題、複選題及填充題等題型來產生「創造」能力的層次試題。

三、「輔助教師產製測驗題庫子系統」與教師出題及現有測驗系統之功能比較

經過電腦輔助產製試題方法的設計及雛形實驗系統的建置後，本研究針對在電子化學習環境中線上測驗系統測驗試題的產製方式，以先導實驗的教師人工出題和本研究提出的電腦產製試題等二種出題方式進行出題效益特性的比較。本研究期望藉由下列的效益比較，期望能更瞭解本研究發展之試題產製方法的優點與限制，詳細的效益比較內容如表 28 所示。

表 28: 電腦產製試題與教師人工出題的效益比較表

出題方式 / 特性項目	教師人工出題	電腦輔助教師產製試題
試題標示出其可用以評量的知識向度及認知向度資訊	先導實驗發現多數教師在發展試題時，並未從知識向度及認知向度來思考，所產製的試題也未標示出其所能評量的向度。若將這些試題直接當作線上測驗系統的題庫，則由於試題未包含這二種向度資訊，因此對於學習者的測驗，除了產生一個測驗成績外，並無法使測驗具有教育意涵上的實質意義。此外，若要標示教師人工出題的試題所屬向度資訊，也需藉由人工的判斷及處理。	本研究藉由教材本體知識及試題類型產製規則，能夠在試題產製完成時，自動將該試題所能評量的知識向度及認知向度資訊記錄在試題中，而不需教師的涉入。而系統也能根據規則，產生不同知識向度及認知向度的試題，當這些試題直接使用在線上測驗系統時，學習者在測驗後，可以了解自己在不同向度上的學習成效，並做為改進學習的參考，因此可使測驗活動能具有實質的教育意義。
試題可用以評量的知識層次與認知層次範圍	本研究在對教師產製的試題進行向度類型分析後發現，教師受限於單選、複選、是非及填充題等題型，產製的試題其所屬的認知向度幾乎仍偏向於記憶、了解、應用及分析等低層次的試題內容。而教師認為評鑑、創造等高層次認知試題，除非有豐富的教學及試題發展經驗，並投入大量的心思去設計題目，否則發展不易。	本研究產生的試題能評量的知識種類包含事實、概念與程序知識，後設知識則難以透過線上測驗方式來評量。而本研究經由試題產製規則的設計，所能產生用以評量記憶、了解、應用、分析、評鑑等五種認知層次的試題，目前僅能標示出 Bloom 的第一層類別概念，而第二層子類別的試題僅能完成部分類別概念的產製與標示。
試題對於測驗回饋與學習者的學習幫助	教師產製試題時，少數教師會在部分試題中，加註相關的知識回饋內容，但大多數教師則未在試題中設計有關於測驗的知識回饋訊息。因此若是這些試題直接被使用在電子化學習環境下，學習者在進行線上測驗後，僅能得到答對或答錯的簡單回饋，以及成績訊息，學習者無法在測驗後取得充足的知識學習輔助。	本系統產生的試題，都會將原始正確的教材知識及相關知識的網頁連結資訊儲存在試題的回饋訊息欄位中。因此學習者在測驗後便可取得額外的知識學習輔助。此外，學習者透過本研究的試題進行線上測驗，學習者還能各種知識向度及認知向度的學習情形回饋資訊。

表 28: 電腦產製試題與教師人工出題的效益比較表(續 1)

出題方式 特性項目	教師人工出題	電腦輔助教師產製試題
建立測驗題庫所需付出之人力與時間	教師自行建立試題時，通常需要重新瀏覽課程教材內的內容，然後逐段確認要評量的知識概念，再逐字逐段輸入及組合試題的語幹、備選答案、配分、題型、答錯回饋內容等資訊。本研究中教師平均發展一個章節題庫所需的時間為 3.9 小時，15 位教師所發展的試題共為 440 題，其中還有許多試題重複，需額外耗費人工進行過濾處理。因此題庫建立需要投入大量的教師時間及心力才能達成。	透過電腦高速運算的特性，本研究的系統能花費 5 分鐘過濾與參與教師相同的指定章節教材知識內容，並根據教材知識特性與關係，自動建立近 18621 題的試題。對教師而言，經由電腦輔助教師產製試題後，教師僅需事後選取要使用的試題即可，關於試題中包含的回饋內容及其他資訊，電腦均能自行協助完成，因此具有試題產製效率的優勢。此外，教師可將省下來的時間用以發展評量高層次知識及認知的試題。
建立大量且充足題庫的能力	受限於教師體力及人類思考方式限制，若要讓單一教師針對特定章節教材，建立大量不重複的試題有其困難，因此教師們認為若要建立龐大豐富的題庫，則需透過多人共同出題來產生不同觀點的試題。單一教師產製試題，容易產生題庫不足的問題。	本研究透過試題產製規則及教材知識本體結構的設計，能夠讓電腦大量產製出規則及知識本體結構所能描述的所有可能試題組合，這對於需求大量題庫以提供學習者進行自我評量的電子化學習環境而言，系統輔助教師產製試題的方式能具有提供龐大題庫的優勢能力。
改善學生直接記憶答案的能力	從測驗理論及過去的研究皆指出，選擇反應型的試題容易產生學生直接記憶答案的問題，因此受限於單一教師出題的試題數量及試題變化，線上測驗的學習者可能在重複進行幾次自我評量後，便把有限的測驗試題答案都記憶下來，而未必真正瞭解每個題目的知識概念。	電腦產製試題的設計允許相同的知識概念能產生很多組試題，因此當系統要重複評量相同知識概念時，學生每次看到的題目都可能不同，因此降低學習者直接記憶每個題幹答案的動機，而促使學習者真正去理解試題所表達的意涵。
題目重複或偏重某概念的情形	若是由教師一個人建立試題，因為教師可以自行控制試題內容，因此比較不會產生題目重複及偏重某概念的情形；但若由多人同時出題，題目重複的情形將無可避免。在實驗中，15 位教師所產製的 440 題試題中，許多試題都是重複的，教師們也普遍認為不同教師的試題重點差異性不高，因此某概念試題會有較多的情形。	本研究的系統能產生許多評量相同知識概念，但是試題間的題型、題幹內容及備選答案並不同的試題。為了避免同一次測驗中出現相同概念的試題，本研究在每個試題上都標記了試題所屬知識概念的編號，以作為系統在進行測驗出題時的識別。此外，本研究的試題產製並非以知識的重要性來選擇，而是以知識本體及規則符合度來決定。因此對於各種知識內容而言，都有機會被產製成具有特定評量意義的試題，因此本研究所產生的試題並不會偏重於某個知識概念。
測驗題型種類	因為配合線上測驗系統的評分能力與限制，教師通在建立線上測驗的題庫時，仍以是非、單選及複選題為主，但有教師指出填充題的答案會有可替代詞彙的問題，一般線上測驗系統直接使用填充題的題庫時，是否能辨識評分是一項挑戰。	本研究因為結合了楊亨利與應鳴雄(民 95)的智慧評分機制，因此除了是非題、單選題、複選題之外，也能產生多空格的填充題，並且在產生填充題時，也會額外在試題中加入評分時評分類型參數資料。對於填充題答案的可替代詞彙問題，系統已針對特定知識領域建立龐大的詞彙關係資料庫，並能將系統產製的填充試題直接應用。

表 28: 電腦產製試題與教師人工出題的效益比較表(續 2)

出題方式 / 特性項目	教師人工出題	電腦輔助教師產製試題
選擇及編製試題的客觀程度	教師通常以個人主觀認知，依照自己認為的知識重點來選擇要編製成試題的知識對象與內容，因此試題會較主觀。	本研究是以試題產製規則來出題，因此沒有人為主觀的出題問題，每個知識概念都能由系統客觀的分析比對，以進行試題產製。
產生結合多種知識的新概念試題能力	少數具有豐富開發試題經驗的教師，能夠發展出整合多個知識概念，或是發展更高層次的試題。但是多數教師所發展的試題，均受限於發展試題時的教材內容，而發展與現存知識內容相似的概念試題。	本系統能在新知識教材概念被加入到教材知識庫之後，重新分析新知識與過去知識的關係。透過知識本體結構，或許可以發現原本分屬不同章節的兩個獨立知識間，事實上存在著某些知識的關係，這個新關係可透過試題產製規則，自動產生新的試題概念。
試題品質	多數的專業系科教師並沒有修習教育理論與試題發展的課程，也沒有教育學分，因此所發展的試題通常違反了特定題型的試題發展規則，而會影響試題品質。但另一方面，有經驗的教師則反而可以透過精心設計的試題，評量出高層次的認知，而產製高品質的試題。而試題鑑別能力品質，則需經過多次測驗後才能客觀計算。	由於電腦產製試題時，已將各題型的重要產製規則建置在系統內，因此試題大致上均能符合試題發展的原則。然而，由於系統產製的試題是依靠產製規則來完成，但是仍有可能產生一些品質不佳的試題，因此本系統提供教師試題審核的功能，以確保系統內的試題能具備一定的試題品質。

第二節 線上測驗與智慧評分子系統之系統成效評估

由於目前現有的線上測驗系統大多已具備基本的出題及評分功能，但卻無法有效針對可能造成評分效力影響的填充題型，進行技術上的探討與解決，因此本研究將以線上測驗支援填充題型的測驗及評分功能，以及評分成績相較於傳統測驗的結果差異進行分析研究，所以在系統的功能績效部份，則以本研究之目的進行選用。經由第二章文獻探討中實際針對國內知名線上測驗系統功能比較結果，以及先導實驗中的研究結論，現有的測驗系統若包含填充題型時，將會引發評分效力降低的問題，其真正的問題在於現有測驗系統的評分機制可能因為不具有語意分析及模擬教師評分行為的能力，導致測驗系統無法具備與傳統測驗相同的評分效力。因此，本研究在進行系統成效評估時，選用的系統關鍵績效指標將包括支援填充題的功能與數量、填充題答案的語意分析能力、消弭引發效力降低問題的能力、線上申訴處理功能、語意詞彙擴充能力、電子試卷保存能力、回饋時機設定能力、模擬教師個人評分特質能力、評分效力、多位教師評分認知衝突解決能力。

本節將針對「線上測驗與智慧評分子系統」的系統成效內容進行描述，此部分的系統評估將從「智慧評分機制之評分效力實證研究」及「與現有測驗系統進行功能特性比較」等二個部分進行。

一、 智慧評分機制評分效力實證研究之研究設計

(一) 問題陳述與假設：本研究允許教師將個人評分之規則、風格特質，以參數設定方式建立至上述評分機制內，期待該評分機制技術能具備與教師紙筆相同之評分效力。不過，由於每個學生皆來自不同的成長環境，所習慣使用的中文詞彙均可能有差異，再加上中文詞彙間具有相似語意者眾多，管理者或教師可能無法在系統

運作初期便在智慧型評分機制中將所有詞彙間的語意關係都考慮周全，不過若系統評分錯誤，應有學習功能對這些新的語意知識予以擴充，以確保在下一次測驗評分中做出正確的判斷。因此，本研究期待下列假設驗證能夠成立。

- H1: 不同的評分機制在測驗成績的評分結果上會有顯著差異。
  - H1a:使用一般型評分機制進行包含填充題型的測驗評分,與紙筆測驗的評分結果會有顯著差異。
  - H1b:使用智慧型評分機制進行包含填充題型的測驗評分，與紙筆測驗的評分結果沒有顯著差異。
  - H1c: 使用智慧型評分機制進行包含填充題型的測驗評分，與一般型評分機制的評分結果間會有顯著差異。而且智慧型評分與紙筆測驗評分間的成績差距,明顯會比一般型評分機制與紙筆測驗評分間的成績差距還小。
- H2: 智慧型評分機制經過詞彙語意的知識擴充後，其評分結果應與未擴充前有差異，而且應更加縮小與紙筆評分間成績差距。

(二) 研究樣本：本研究以本研究以中部某技術學院資管系學生 3 班 120 位修管理資訊系統課程之同學為樣本，採實地實驗法進行，以探討線上測驗系統環境中，評分機制對於測驗成績的影響。

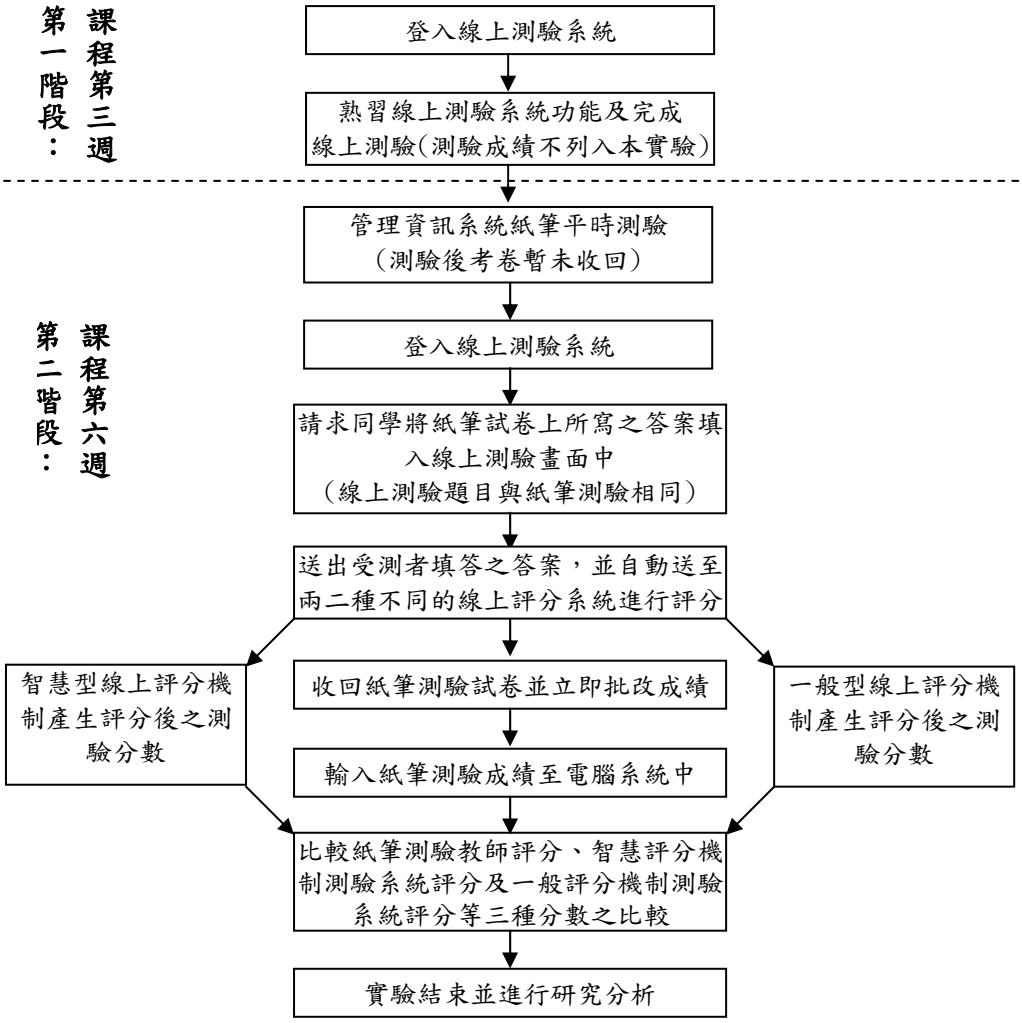

圖 67：評分效力研究之實驗流程圖

(三) 資料收集方法：由學生實際參與紙筆測驗及線上測驗，並取得經由不同評分機制的評分成績結果來進行資料分析。

(四) 實驗流程：本實驗在進行之前先根據表 1 的國內知名線上測驗系統之功能，依據其評分機制概念設計出「一般性評分機制」之評分模組，另一方面再依據本研究提出之技術方法，設計出「智慧型評分機制」之評分模組。為了使兩種不同評分機制能夠進行評分效力的比較，因此這兩種評分機制共用相同之測驗介面，受測學生在單一測驗介面中填答問題答案後，系統會將試卷送至這兩種不同的評分機制中進行評分，並分別計算出評分後之成績結果。本研究之實驗流程如圖 67 所示，共分成二個階段實施。第一階段的目的是為了使受測學生能熟悉測驗系統之操作功能，並降低受測學生因不熟悉系統操作而產生實驗干擾。此階段刻意選擇第三週進行正式授課後的第一次線上平時測驗，受測學生皆於上課時透過電腦進行線上測驗，並於線上填寫電腦網路使用之基本資料，但是本次之成績結果則不做為本研究之分析。第二階段則於課程第六週進行管理資訊系統課程第二次平時測驗，教師首先在智慧型評分機制中，依照教師批改填充題的規則完成相關參數設定，隨後教師發給每位同學一份紙筆測驗的試卷，內容共計 17 題，非填充題題型包括 5 題是非題、5 題單選題及 2 題複選題，每題 5 分共計 60 分；填充題型有 5 題，共含 8 個填充格，共計 40 分。為控制測驗時間對測驗成績的影響，本實驗進行時並未告知學生實際考試的時間限制，並期望每位同學均能有充分時間作答(考題設計時以 30 分鐘內可填答完成為標準進行設計)。當所有所有學生均完成紙筆測驗的填寫後，此時教師則要求受測同學登入線上測驗系統開啟線上測驗的試題，此份線上試卷之試題與同學手上之紙筆測驗試卷內容完全相同，教師要求同學將自己在紙筆試卷上所寫的答案，按照題目順序將相同題目之答案照實的輸入至線上試卷中，並當答案全部輸入完成後，則送出試卷完成測驗。當所有學生都完成線上測驗的程序後，教師向受測同學收回所有紙筆試卷，並由教師親自批改，批改後的成績再輸入至資料庫中儲存。另一方面，當測驗系統收到學生送出之答案後，系統會自動將受測者之答案分別送至「智慧型線上評分機制」與「一般型線上評分機制」進行評分，不同評分機制所產生之評分成績結果則儲存至資料庫中。因此，每位受測者完成此實驗後，同一份試題會出現來自三種不同評分機制所產生的成績，這些成績資料將成為後續研究的分析資料來源。

(五) 研究架構與變數說明：圖 68 為此實證研究之研究架構，透過不同的評分機制來分析對於測驗評分成績的影響，研究架構中的變數說明如下。

1. 自變數：僅有「評分機制」一項，包括三種評分機制，分別是一般型評分機制、智慧型評分機制及紙筆評分。
2. 依變數：Bostorm(1990)及許多探討測驗成效的研究中，經常會使用測驗成績來當作評量學習成效的指標。Bugbee(1996)也認為不同的評量工具若能具有相同的評分結果，則具有評分效力。因此本研究直接採用線上測驗的評分成績結果與教師親自批改評分的成績結果進行比較，以觀察評分工具對於測驗成績的影響是否有差異。

3. 控制變數：為使測驗時間不會去影響測驗結果及評分效力，本研究在實驗進行時，採取寬裕測驗時間。另外，為隔絕介面變數干擾，本研究之一般型與智慧型評分採相同介面，並由系統將收到之電子試卷，自動分派至不同評分機制。

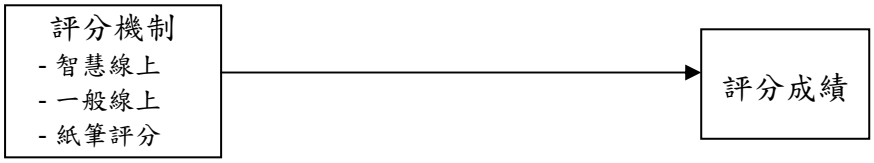

圖 68：評分效力研究之研究架構圖

二、 智慧評分機制評分效力實證研究之資料分析

在此部分之實證研究方面，本研究採用 SPSS 統計軟體作為資料分析的工具軟體，分析之結果說明如下。

(一) 樣本基本資料分析

此部分之實證研究共有 120 位受測者參與實驗，其中男生 63 位，女生 57 位。但經過第一階段及第二階段實驗後，由於有 8 位受測者未全程參與實驗或錯誤操作系統導致其實驗資料未能完整取得，因此後續資料分析將扣除此 8 位受測者資料，因此全程參與之受測者共有 112 位，其中男生 58 位，女生 54 位。

(二) 研究假設檢定

此部分之實證研究最終共有 112 個受測者，每位學生會先進行紙筆測驗，再進行線上測驗，最後會得到三種評分工具的評分成績，此成績為包含所有題型之成績，結果如表 7 所示。受測者在紙筆測驗教師評分的平均成績為 53.31 分，而智慧線上測驗評分機制成績為 52.21 分，一般線上測驗評分機制成績為 48.81 分，從分數上可看出智慧型評分成績與教師親自批改的紙筆測驗成績差異較小。為進行本研究之第一項假設檢定，本研究使用林清山(1990)多變項分析統計法中的相依樣本單一組重複量數統計分析方法，以瞭解受測者對於不同的評分機制所得到的測驗成績評分結果是否有顯著差異。由於使用重複量數的概念來進行，因此不能將測驗成績直接進行分析，而需利用每位受測者在各種評分機制所得到之成績，分別計算出其受測者的紙筆與一般評分成績差距(使用 PG 符號表示)、紙筆與智慧評分成績差距 (使用 PI 符號表示)、智慧與一般評分成績分數差距(使用 IG 符號表示)等三項資料，此三項資料之平均結果如表 29 所示，其中紙筆評分與一般評分成績之分數平均差距高達 4.5 分，顯示這兩種評分機制的評分結果差異較大;另外，紙筆評分與智慧評分成績之分數平均差距最小，但仍有 1.10 分的差距。

表 29 受測樣本之測驗結果

	項目	樣本數	平均分數	標準差
原始資料	一般 OLT 評分之成績	112	48.81	16.06
	智慧 OLT 評分之成績	112	52.21	16.55
	紙筆教師評分之成績	112	53.31	16.87
資料轉換	紙筆評分與一般評分成績之分數差距(PG)	112	4.50	4.52
	紙筆評分與智慧評分成績之分數差距(PI)	112	1.10	2.40
	智慧評分與一般評分成績之分數差距(IG)	112	3.40	3.57

經過 SPSS 的相依樣本單因子多變量變異數分析(考張紹勳，1997)，進行不同評分機

制評分成績的多變量顯著性檢驗，結果如表 30 所示，其中 Wilks Λ 值為 0.490，P<0.05，表示受測者在不同評分工具所得到的成績結果並不相同，因此支持了本研究所提出的 H1 假設，不同的評分機制在測驗成績的評分結果上會有顯著差異。

表 30 評分工具假設之變異數分析表

檢驗項目	Wilks' Λ值	F值	P值	Eta^2	顯著性α=0.05
評分工具	0.490	F(2,110)=57.293	0.000	0.510	達顯著水準

表31為評分工具間之評分成績差異檢定分析結果，無論是「紙筆評分與一般評分(PG)」、「紙筆評分與智慧評分(PI)」及「智慧評分與一般評分(IG)」等工具間的成績差異檢定，P值均小於0.05，均達到顯著水準，顯示H1a、H1c假設均獲得支持：受測者在紙筆與一般評分機制之成績有顯著差異、智慧與紙筆評分間差距小於一般與紙筆評分間差距。但是，受測者在紙筆與智慧評分機制之成績上也有顯著差異，因此H1b假設並未獲得支持。

表31 評分工具間之評分成績差異檢定分析表

相依變數	平均數	標準差	t值	P值	顯著性α=0.05
PG	4.50	4.52	10.531	0.000	達顯著水準
PI	1.10	2.40	4.848	0.000	達顯著水準
IG	3.40	3.57	10.098	0.000	達顯著水準

為了確認造成評分工具間之評分成績差異是否源自於填充題型產生的評分差異，因此本研究另外將填充題型及非填充題型的評分成績結果分別重新進行評分工具間之評分成績差異檢定。

在僅包含填充題型評分成績結果中，一般 OLT 評分之平均成績為 17.58 分(S.D=10.04)，智慧 OLT 的評分平均為 20.81 分(S.D=10.29)，紙筆教師評分之平均成績為 21.82(S.D=10.39)。在僅計算填充題型的測驗成績時，「紙筆評分與一般評分(PG)」、「紙筆評分與智慧評分(PI)」及「智慧評分與一般評分(IG)」等工具間的成績差距分別是 4.41 分、1.01 分、3.40 分。在不同評分機制評分成績的多變量顯著性檢驗上，Wilks Λ 值為 0.515，P 值=0.000，顯示出受測者在不同評分工具所得到的填充題型成績結果並不相同。而評分工具間之評分成績差異檢定分析結果，無論是「紙筆評分與一般評分(PG)」、「紙筆評分與智慧評分(PI)」及「智慧評分與一般評分(IG)」等工具間的成績差異檢定，P 值均小於 0.05，均達到顯著水準，其統計值與統計結果與計算所有題型成績的表 9 結果相當類似。

在僅非填充題型評分成績結果中，一般 OLT 評分與智慧 OLT 之平均成績均為 31.31 分(S.D=8.98)，紙筆教師評分之平均成績為 31.40(S.D=9.14)。在僅計算非填充題型的測驗成績時，「紙筆評分與一般評分(PG)」、「紙筆評分與智慧評分(PI)」及「智慧評分與一般評分(IG)」等工具間的成績差距分別是 0.089 分、0.089 分、0.000 分。在不同評分機制評分成績的多變量顯著性檢驗上，Wilks Λ 值為 0.991，P 值=0.319，顯示出受測者在不同評分工具所得到的非填充題型成績結果並無顯著差異。

由上述分析結果發現，不同的評分機制在非填充題的評分結果上並沒有顯著差異，線上測驗所產生的評分結果差異幾乎都來自於填充題型的評分結果，因此使用整體分數進行評分成績結果的差異檢定時，所得到的分析結果會與僅考慮填充題型評分成績時的結果極為類似，因此本研究仍以考慮各種題型的整體分數進行分析。

本研究原先認為電腦中若使用智慧型評分機制，由於該機制已儲存了教師的批改規

則及習慣，並藉由語意詞彙資料庫的建立，建立了測驗知識領域的語意詞彙間關係，因此應能與教師親自評分的結果相似。但經過上述實驗及統計分析後，雖然使用智慧評分與紙筆評分的成績平均只有差距 1.10 分，但仍舊在統計上顯示無法有相同的評分效力。仔細分析形成原因，可能來自於學生填答的填充題答案與教師設定的填充題標準答案不同，雖其語意詞彙與標準答案間存在同義或相似關係，只因為系統語意資料庫所儲存的語意關係知識仍舊不足，才會形成智慧評分無法與紙筆評分有相同的結果。事實上，在第六週測驗完成後，部分學生的確也在測驗系統中針對某些填答的答案提出成績申訴，因此，本研究便提出了第二項假設，認為只要智慧型評分機制經過詞彙語意的知識擴充後，其評分結果會比未擴充前更加提升。

為了進行此項假設檢驗，本研究針對學生所申訴的問題進行處理，並對於系統內建語意詞彙關係不足的知識部分進行擴充。學生所申訴的內容主要是針對答案詞彙語意的部分，例如某填充題之題目為「從作業系統的處理方式而言，當資料收集到一定時間或一定量才處理，稱為___。而將 CPU 時間平均分配給每個使用者程式的作業系統處理方式稱為___。」，本題在系統內的標準答案分別是「批次處理」與「分時處理」。以「批次處理」詞彙而言，「批次處理」在系統中已存在的相關語意詞彙包括「batch processing」、「Batch」、「整批處理」、「批次」等 6 個詞彙，但是有一些學生所寫的答案是當初教師及系統未想像到的，諸如「批次作業」、「批次作業處理」及「整批處理作業」，這些答案在教師紙筆評分時被認為是與標準答案相同，但在智慧型線上評分機制中卻無法對這些相同語意的詞彙進行辨識，導致評分結果產生差異。另外，像是在另一個填充題題目為「___能將企業的智慧資產透過資訊科技累積起來，並能進行有效的運用，並達到企業員工間能快速傳遞及分享知識經驗，使企業能不斷創新。」中，系統預設之標準答案為「知識管理」，系統內建之不同程度之相關語意包括「知識管理系統」、「知識庫系統」、「Knowledge Management」等，而學生則輸入了「知識系統」答案，因為教師在系統初期語意建立時並未將「知識系統」的詞彙建立到語意詞彙關係資料庫中，因此系統也無法做出正確的判斷評分，而在其他題目中也均有類似的情形發生。由於本研究設計之測驗系統保留了受測學生當初所填寫的電子試卷資料，本研究在完成學生申訴的問題及擴充所遺漏的相關語意詞彙後，將針對這些受測者當初所填寫的電子試卷答案，重新送至智慧評分機制重新評分，再與未擴充詞彙語意前之智慧評分結果、紙筆評分結果進行比較。

本研究針對「未擴充詞彙語意前之智慧評分」、「擴充詞彙語意後之智慧評分」及「紙筆評分」進行不同評分機制評分成績的多變量顯著性檢驗，結果如表 32 所示，其中 Wilks Λ 值為 0.821，$P<0.05$，表示「未擴充詞彙語意前之智慧評分」、「擴充詞彙語意後之智慧評分」及「紙筆評分」等不同評分工具所得到的成績結果並不相同，因此仍支持本研究所提出的 H1 假設，不同的評分機制在測驗成績的評分結果上會有顯著差異。

表 32 包含擴充詞彙語意後之智慧評分機制的變異數分析表

檢驗項目	Wilks' Λ值	F值	P值	顯著性α=0.05
評分工具	0.821	F(2,110)=12.023	0.000	達顯著水準

表33為「擴充詞彙語意後之智慧評分」與「未擴充詞彙語意前之智慧評分」及「紙筆評分」評分工具間之評分成績差異檢定分析結果。表中顯示「紙筆評分與已擴充語意詞彙後的智慧評分(PI2)」間的成績差異未達顯著水準，表示經過擴充語意後的智慧與紙筆評分的成績並無顯著差異，具有與紙筆評分相同的評分效力。因此，若是線上測驗系

統的評分機制經過學生實際測驗後的語意擴充，應能使得智慧評分機制具有與紙筆測驗相同的評分效力，先前未得到支持的H1b假設，在擴充語意詞彙後而得到支持。

另外，表33中也顯示「已擴充語意詞彙後的智慧評分」與「未擴充語意詞彙前的智慧評分」間的評分結果有顯著差異，原本未擴充前與紙筆的平均評分差異是1.03分，而已擴充後與紙筆評分的平均評分差異則縮減為0.071分，顯示智慧評分機制經過不斷的擴充語意詞彙後，將能提升其評分效力，因此假設H2獲得支持。

表33 評分工具間之評分成績差異檢定分析表

相依變數	平均數	標準差	t值	P值	顯著性α=0.05
PI2	0.071	0.07	1.02	0.312	
II2	1.03	0.21	4.83	0.000	達顯著水準

註： PI2 表示「紙筆與已擴充語意詞彙後的智慧評分」間之成績差異
II2 表示「未擴充與已擴充語意詞彙後的智慧評分」間成績差異

雖然經過語意擴充後的智慧評分機制能具有與紙筆測驗相同的評分效力，但是這兩種評分機制的成績結果平均仍有 0.071 分的差異，本研究再將這些差異的受測資料進行比對發現，這些差異是屬於實驗過程中受測者的非刻意之隨機誤差，其中包括受測者在將紙筆測驗上的答案抄錄至線上測驗系統時，少填答了一題答案，或電腦輸入時拼錯字及漏字，而產生紙筆測驗的答案與線上測驗輸入的答案並不一致，因此導致仍出現成績差異，若排除受測者從紙筆測驗抄寫到線上測驗過程中的人為隨機誤差，則利用智慧型評分機制來取代紙筆測驗時的評分效力將更能提升。

(三) 智慧評分機制之評分效力實證研究之結果論述

線上測驗系統若只處理是非題、單選題、複選題等具有固定答案的測驗題型時，並不會產生與紙筆測驗結果不同的測驗結果。但若是要進一步推展線上測驗系統的測驗題型應用範圍，並於測驗中加入填充題型等具有眾多可能的相同或相似語意詞彙答案時，則需要注意其評分機制的評分效力問題。本研究在探討各種評分機制的評分效力，主要是以各評量工具相對於教師親自批改評分之成績結果差異，做為評分效力的依據。一個完美的電腦評分機制應該與教師親自批改的評分成績結果相同，因此一個評分機制工具的評分結果與紙筆評分的結果越接近，或是在統計檢定上呈現出沒有顯著差異，則表示該評分機制具有與紙筆測驗相同或相似的評分效力。

綜合資料分析結果，在包含填充題型的測驗中，不同的評分機制在測驗成績的評分結果上會有顯著差異，智慧型評分與紙筆測驗評分間的成績差距，會明顯比一般型評分機制與紙筆測驗評分間的成績差距還小。但「未擴充語意前智慧評分機制」的評分效力無法與紙筆測驗等化，若經過學生實際測驗後的語意擴充，將能使得智慧評分機制具有與紙筆測驗相同的評分效力。

不過，即使本研究證明智慧型評分機制經過語意詞彙關係的知識擴充後，將可具備與紙筆評分相同的評分效力，並不表示線上測驗系統透過智慧型評分機制所評分後的成績結果會與紙筆評分的結果 100%相同，這二種評分機制的成績結果間可能只能近似100%而已。歸納其原因，智慧評分機制內可能永遠無法將全世界中所有詞彙與詞彙間具有相同或相似語意的關係知識全部建立完整，教師也無法將標準答案相對應的所有相同或相似詞彙事先完全沒有遺漏的建立至系統內，學生也可能因為來自不同的背景，而有許多未預期的答案寫法。除此之外，學生在透過線上測驗系統進行測驗時，也可能遭受

到電腦環境與技術的諸多干擾或其他隨機性的因素，而導致線上測驗的結果與紙筆測驗結果產生差異，其中包括對於電腦經驗、電腦焦慮程度、中文輸入法的使用、非刻意遺漏答案部分字元等。

根據本研究結論，對於用線上測驗系統做為評分工具的教師及從事線上測驗系統發展工作者而言，可以持續透過智慧評分機制、語意詞彙關係的發展及教師評分規則參數的建立，使得線上測驗系統提供填充題題型時，仍然可具備與紙筆測驗相同的評分效力。然而本研究未來仍需透過更多不同對象及測驗科目的評分結果分析，將智慧型評分機制的成效推論至其他的對象領域及科目範圍。此外，未來對於問答題測驗題型的評分效力研究，則還需要透過中文詞句語法結構的分析來進一步克服。

三、「線上測驗與智慧評分子系統」與現有測驗系統之功能比較

本系統經過建置完成後，本研究將此系統之諸多功能特性與其他現有的測驗系統進行比較，以期能凸顯本研究發展之測驗系統具有更佳的測驗效力。其相關之分析內容如表 34。

表 34: 線上測驗與智慧評分子系統相較於現有系統的功能特性分析表

功能特性	本系統的成效說明
支援建構型題型之多格填充測驗	目前的測驗系統大多數都僅支援選擇反應型的測驗類型題目，諸如是非題、選擇題及複選題。縱使有少數系統提供填充題測驗，但也僅能提供 1 個格子的測驗，且不具有智慧型的語意分析機制，本系統每一題填充題最多可建立 20 個填充格。
語意分析功能	本系統提供之填充題測驗，具有語意評分之功能。每個填充題的試題，可共用系統已建立的語意詞彙或使用個人化的相似語意詞庫，此外，教師亦可針對特定試卷中特定題目，自訂該題專用的語意字彙；如果該題答案較為嚴謹，不接受任何相關語意的答案，亦可透過功能關閉該題的語意評分功能。因此本系統的語意評分機制，提供整體系統共用相似語意詞庫、教師個人化詞庫及單一題目專用詞彙組合等三種不同的語意分析層次。
集合順序功能	由於現有測驗系統並未特別針對多格填充答案間的集合順序評分功能進行設計，本系統提供此功能，使得測驗者可如同紙筆測驗般，針對集合順序可互換的填充題目進行填答，而不需擔心順序填錯系統評分錯誤的問題。
提供學生成績申訴管道及申訴處理即時通知功能	當學生對系統評分結果提出質疑或認為評分不合理時，本系統提供學生線上立即申訴的功能，當系統收到申訴案件時會立即透過本系統之電子郵件自動發送元件立即將申訴案件透過 Email 告知出題教師處理，當教師完成申訴處理案件後，系統也會立即將申訴結果透過 Email 告知申訴人。此外，申訴人在申訴期間也可於線上查看申訴案件的處理狀態與進度。
藉由申訴功能達成相似語意詞彙擴充	由於系統很難將世界上所有詞彙的相似語意全部建立至語意庫中，因此學生若在申訴案件中認為系統評分時無法辨識的語意而造成評分錯誤，教師則可藉由此功能來改善評分效力，並不斷擴充語意詞彙內容。
電子試卷的保存	現有諸多測驗系統通常僅將學生的測驗成績結果記錄下來，並未將學生測驗的答題內容及測驗歷程資訊保存下來，若遭遇到學生對測驗結果的爭議，將產生查無當時試卷答題內容的窘況，本系統透過電子試卷的保存功能，允許教師及學生調閱歷史測驗試卷資料。
消弭透過電腦媒介輸入時的誤差	由於某些測驗者在測驗過程中，不小心按到全形功能或按到字母大寫功能，造成評分錯誤，本系統則透過評分機制的改進，消弭類似的評分誤差。

表 34: 線上測驗與智慧評分子系統相較於現有系統的功能特性分析表(續)

功能特性	本系統的成效說明
中文輸入環境的誤差改善	大部分的測驗在測驗學生的知識認知，但卻可能因學生中文能力逐漸降低，而學生卻透過輸入法選出原本不會寫的中文字或因為中文程度而誤選了同音異字的錯字，本系統可由老師自己設定「斟酌」給分的模糊機制，使學生在使用「注音輸入法」選到同音異字的錯字，且不影響整個答案表現時，不置於被評分為全錯。
成績公告時機	由於本系統考慮教師進行測驗的動機，因此提供此項特有功能。由於目前有些測驗系統只是提供學生自我練習，學生可在測驗之後立即得到成績及答案回饋。但有些測驗系統則是採用延遲回饋，該系統必須等開放測驗的時間截止後，所有學生考試完畢，才可查詢成績。本系統提供教師設定每份測驗的成績及答案回饋時機，回饋的時機根據理論從立即回饋、延遲回饋及不給予回饋等概念，共有 11 種回饋時機供教師選擇。如此一來將可避免先考完的學生知道分數及題庫的答案後，洩漏給其他同學，而產生測驗公平性的問題
可接受之漏字評分功能	如同生活中人們也會不經意的漏寫字一般，如果漏寫的字本身不影響答案表現，則通常老師亦會斟酌給分，因此本系統亦將教師的此項特性加入於本系統中。
教師個人評分特質參數設定	為使測驗系統能具有與出題教師相同的評分特質，因此本系統提供評分機制參數供教師設定，教師可依照自己平時對於漏字、同音異字及每個語意詞彙間的關係認知，自行設定相關參數，以期測驗系統能與教師的評分方式一致，以增加線上測驗的測驗評量效力。
重視線上測驗的測驗效力	國內外大多數線上測驗系統被視為是網路上的測量工具，但並未考慮提供的線上測驗系統其測驗效力是否與紙筆測驗等化。本研究則從測驗的基本目的出發，並採用測驗效力觀點發展線上測驗系統，使測驗系統能回歸到測驗效力的本質。
結合系統內所有教師共識，提供高品質的系統評分參數及相似語意詞庫	目前的測驗系統大多數未提供教師設定教師個人化的評分風格參數，而且也未使用模糊理論技術來分析與結合系統內所有教師的共識，產生高品質的系統內建評分參數及相似語意詞庫相似值，隨著系統內教師使用人數的增加，系統的評分參數值及相似語意詞庫相似值將更趨於穩定與合理。
解決多位教師在評分認知上的衝突	本研究可提供教師個人化的評分風格參數設定，以保留教師個人的評分風格，同時也能藉由系統內建且自動根據系統內所有教師的認知，產生具有大多數教師共識的評分參數及評分風格，因此可藉由具備多數教師共事的系統評分風格，來進多位不同授課教師的全校性統一測驗評分。