

虛擬測驗服務中心-- 建構在 Internet 的適性測驗環境*

何榮桂 陳麗如 郭再興 蘇建誠 籃玉如

國立臺灣師範大學資訊教育研究所

臺北市和平東路一段 162 號

TEL:(02)3944288 EXT. 57

EMAIL: s13095@cc.ntnu.edu.tw

摘要

本研究旨在運用項目反應理論、電腦化適性測驗原理與網路技術，建置一個在 Internet 上之虛擬測驗服務中心，提供題庫管理、線上測驗、線上練習等功能，以期開創一個新的遠距測驗實施模式，冀期(1) 結合教師及測驗專家的專業知能，以建立 豐富、多元、共享的高品質題庫；(2) 提供多樣化的測驗實施模式，給予使用者更 精確、適切的教學評量工具；(3) 突破時空限制，給予使用者線上練習的機會；(4) 提供一個便於自我探索、瞭解自我的途徑；以及(5) 達成測驗的評估、預測、診斷 等功能。

關鍵字：項目反應理論、電腦化適性測驗、線上測驗、線上練習、題庫

*本文係國科會科教處研究計畫「網路環境題庫與測驗之整合系統」(NSC86-2511-S-003-017-CL)之一部分。

壹、緒論

一、設計動機與目的

測驗在人類社會中佔十分重要的地位。舉凡國家之文官甄選、證照頒給、升學甄試、企業之用人選才以及教育上之學習評量、診斷、預測、輔導等應用，皆需使用到測驗。利用無遠弗屆的網際網路(Internet)來建立測驗服務中心，不僅可滿足社會大眾對測驗、練習及自我瞭解的需求，更可充分發揮網路資源共享之特性，實為最具經濟效益之應用。

網路科技的進步和項目反應理論(item response theory ; IRT)及電腦化適性測驗 (computerized adaptive testing; CAT)理論的成熟，促使虛擬測驗服務中心(virtual testing service center)的理念得以實現。建立虛擬測驗服務中心，可以整合教師及測驗專家的專業知能，建立豐富、多元、共享的高品質題庫，提供給使用者更精確、適切的練習、測驗、自我學習評量以及自我探索的測量工具，以達成測驗的評估、預測、診斷等功能。

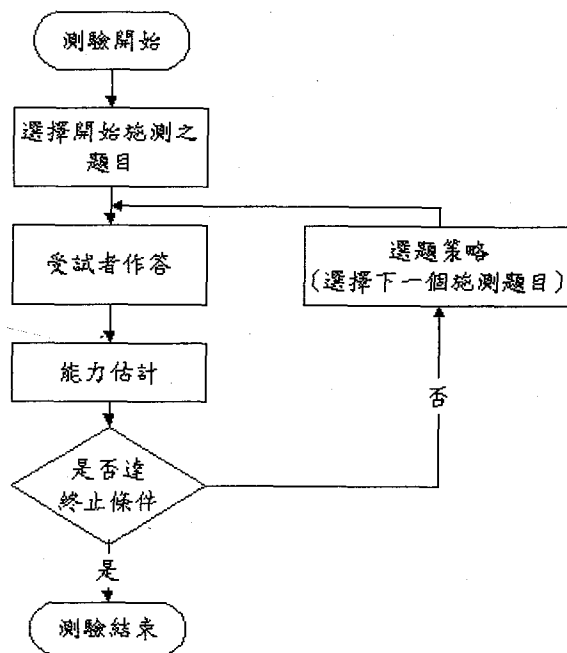
二、範圍與限制

- 1.本研究旨在建立一個虛擬測驗中心，因此著重運用既有的理論（如IRT,CAT等）於題庫管理、線上練習、線上測驗等功能的實作。
- 2.本研究中所稱之測驗乃依測驗的反應型態，區分為最大表現(maximum performance)(如成就測驗、智力測驗、性向測驗等)和典型表現(typical performance)(如人格、態度、興趣量表等)測驗等兩大類，在實施模式方面則包含適性及非適性兩種施測方式。

貳、文獻探討

一、電腦化適性測驗

電腦化適性測驗之實施係根據項目反應理論的原理，程序如圖一，即在施測時，每次皆依受試者之反應，運用電腦快速計算的特點，不斷的重新估計考生能力，再依據選題策略，自題庫中選取最能符合受試者能力的試題施測，直到受試者的能力可以被精確的估計為止。其優點為只需較少的題目，即可精確估計受試者能力，因為估計是在相同的尺度(common scale)上，因此不同受試者之測驗結果仍可相互比較。



圖一 適性測驗實施流程

因本研究所應用的測驗包含最大表現和典型表現兩大類，而此兩類測驗所使用IRT估計模式略有差異，茲分別敘述於下：

1.最大表現測驗採三參數之兩點計分IRT模式

此模式主要建立在兩個基本概念之上：(1)受試者在某一測驗試題上的表現，可以由一組(或至少一個)因素來預測或解釋，稱之為潛在特質(latent traits)或能力(abilities)，(2)受試者的表現情形(項目反應)與潛在特質間的關係，可以由一個連續的試題特徵函數(item characteristic function)來表示，其

圖形稱為試題特徵曲線(item characteristic curve)，可以函數來表示之。根據過去的研究，採用三參數(a,b,c 分別表示鑑別度、難度及猜測度參數)IRT 模式之 CAT，最能收適性的效率(Ho,1989)。

2. 典型表現測驗採多點計分之 IRT 模式

多點計分模式是由兩點計分模式擴充而來，由於本研究之典型表現測驗以李克氏量表為主，因此採選項間具次序關係之等級反應(graded response; GR)模式。此模式是由雙參數模式擴充來，考慮的試題參數包括試題鑑別度、試題難度和選項難度。此模式中，每位受試者對試題之反應均為一連續變數，因此在各能力階層上，都存在受試者反應的常態條件分配，這些條件分配的曲線會在反應的邊際交錯，藉此可計算出每一反應類別對應至能力水準之機率值，由於反應間具有次序關係，因此若受試者的反應落在兩個邊際間機率值之間，則該選項即可表示受試者的特質或態度的強度(Baker, 1992)。

二、試題資料庫

廣義的題庫是指一群便於使用的題目集合(Millman & Arter, 1984)。此定義視題庫為消極角色，僅為儲存試題的地方，為被動的心理測驗工具。狹義的題庫則是指依試題的內容、類別(如單元或學科)或統計特質(如難度、鑑別度參數等)加以整理、分類，並以數位化方式儲存，採電腦來管理及操作的試題組合(Hsu & Sadock, 1985)。

本研究中所規劃之題庫乃運用電腦科技，使題庫成為積極的教學評量工具，可提供教學者、受試者、輔導者所需資訊的功能，並蒐集有關試題之分析資訊，如試題反應理論之鑑別度、難度、猜測度等參數值、或古典測量理論之試題分類代碼、屬性、題型、難度指標(difficulty index)、鑑別度指標(discrimination index)、二系列相關係數、點二系列相關係數、測驗的信度與效度等，供測驗編製者進行更深入的研究探討。所以在規劃時，古典測驗和 IRT 基礎的適性測驗題庫並存，以因應使用者不同需求的題庫，充分發揮題庫的功能。

題庫管理包括試題資料庫的規劃、建立和維護等工作。本研究所規畫之題庫管理功能除了便於輸入、擴充、搜尋、刪除、更新、修改等管理功能之外，也重視試題編製的品質，以作為支援本研究中所建立之適性及非適性測驗。因題庫多屬中、小型資料庫，所以可採用結構化系統分析與設計方法來規劃、設計、開發(Ramez & Shamkant, 1994)。

三、線上練習 VIPQ 策略

為提高學習者的學習成效，應提供其充分且有效的練習機會。然而，影響練習的品質則有下列幾項因素：(1)練習單元的介紹，(2)練習題之特質，(3)選題策略，(4)回饋，(5)引起學生動機等(Stephen & Stanley, 1991)。筆者之一(何榮桂，民 84)曾以英語同義字或詞為練習內容之題材，比較不同選題策略(串列、隨機、閃爍卡式佇列及變數式題距表現佇列等四種)對學習效果的影響，結果發現不同的選題策略會影響練習效果。其中以「變數式題距表現佇列」(variable interval performance queuing; VIPQ)選題策略之練習效果最佳。VIPQ 策略係指將受試者答錯之試題，以變化之間隔

(variable interval)方式，安插於後續的練習題中，提供簡單的重複練習機會，讓練習者練習其尚未精熟的試題。

參、研究方法與步驟

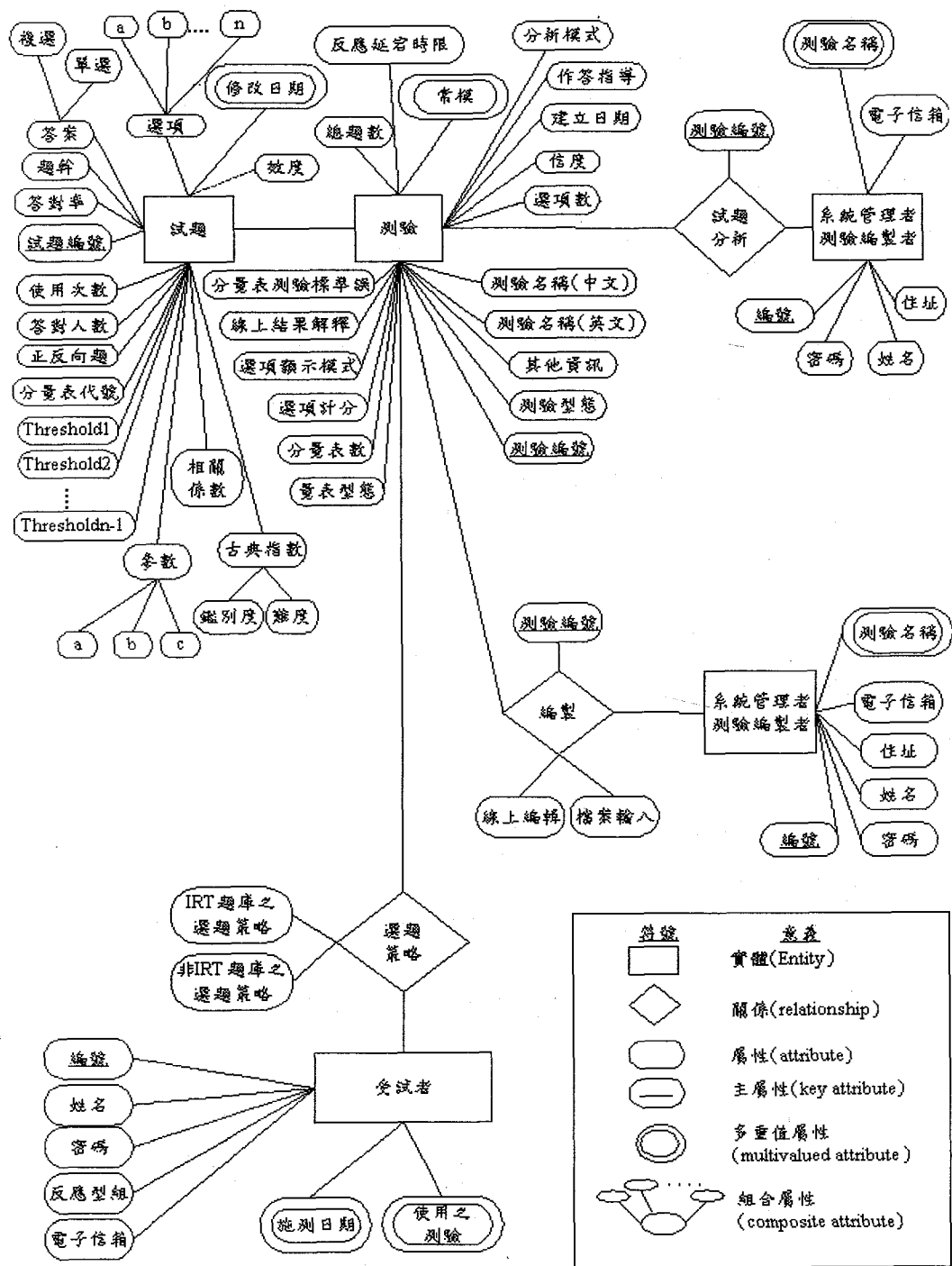
一、設計環境與工具

本研究以 Internet 作為網路骨幹，以 PC 以及 Windows NT server 作為 server 端工作平台。開發語言包括 HTML (hypertext markup language)、JavaScript、Delphi with Paradox Database 等。支援 Netscape 2.0、HTML1.0 以上版本。在 client 端則藉由 HTML 與 JavaScript 及選題模組的配合，以呈現試題或練習題。受試者輸入答案後，透過由 Delphi 所撰寫的 CGI 及線上施測模組，將答題反應傳送回 server 端，在比對答題反應是否正確後，經由選題策略的運作安排下一個施測試題，並自題庫中取出適當的試題傳送回 client 端，讓受者繼續進行測驗或練習，線上施測活動結束為止。

二、系統分析

題庫是測驗服務中心的核心，其品質影響施測結果的精確度，為達成本研究之目標，建置程序使用結構化系統分析設計方法(Ramez & Shamkant, 1994)，即包含需求分析、邏輯設計、實體設計、系統開發等步驟。

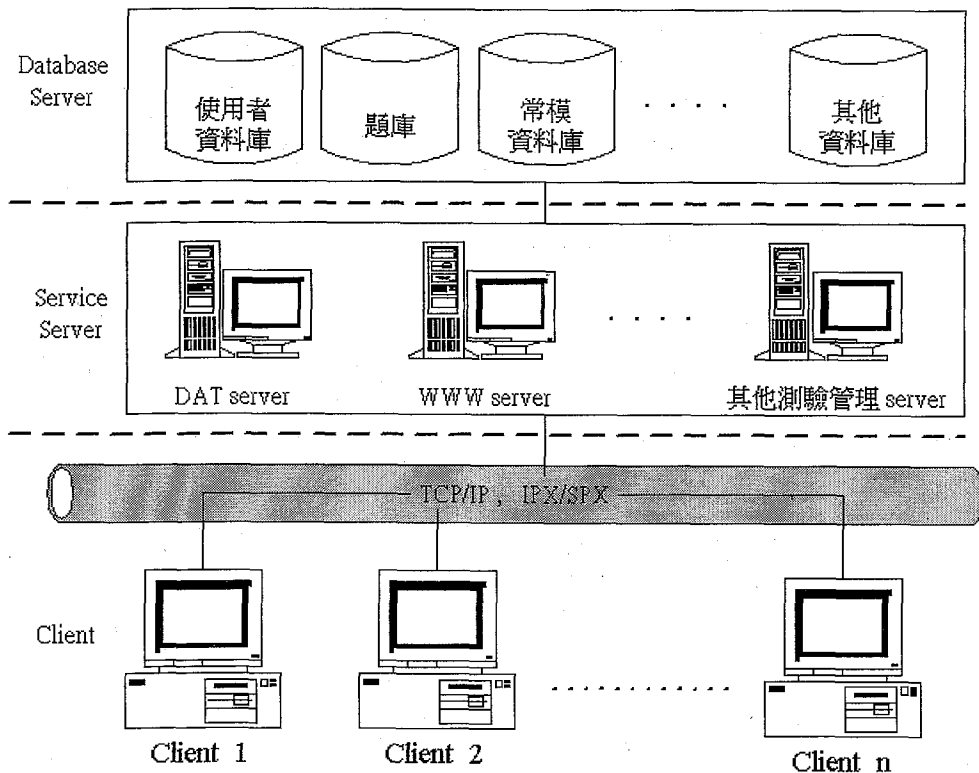
在需求分析階段之主要工作為進行系統之初步規劃工作，本研究中先以實體關係圖 (entity-relationship diagram)來描述系統中各實體單元(entities)之關係(如圖二所示)，系統之功能需求包括：(1)系統中之實體(entities)包含試題、測驗、系統管理者、測驗編製者及受試者等；(2)系統管理者對測驗之關係包括編製(或維護)及進行試題分析；(3)測編製者與測驗之關係包含編製及進行試題分析；以及(4)受試者與測驗之關係則為經選題策略(循序、隨機、VIPQ、最大訊息選題法)接受線上練習或測驗。



圖二 系統之實體關係圖 (ER Diagram)

三、系統設計

虛擬測驗服務中心係建構於 Internet 中, 採用三層式主從遠距測驗系統架構 (three-tier client/server distance testing system architecture) (如圖三所示), 並依前階段之規劃結果, 建立簡單、基本的資料庫綱目(schema)及內部資料結構, 設計各功能需求之合理流程。茲將此階段之設計重點簡述於下:



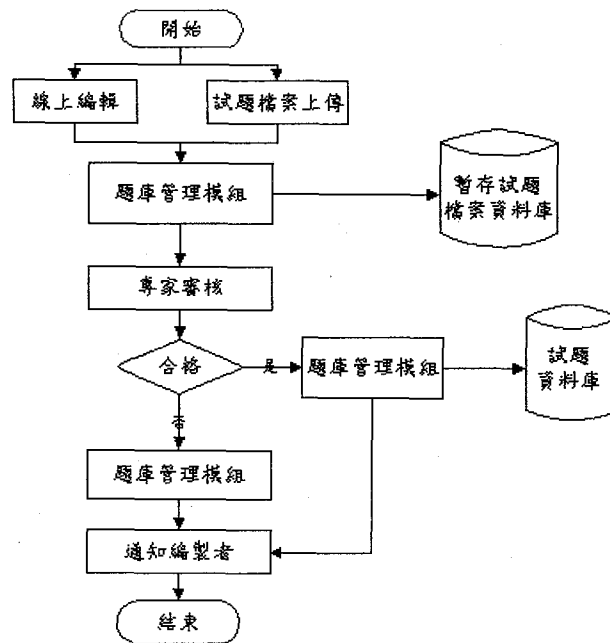
圖三 三層式主從遠距測驗系統架構

1. 三層式主從遠距測驗系統架構

在系統中需建立使用者、試題、常模等資料庫，各資料庫之管理維護工作由資料庫伺服器 (database server) 來負責。為達成虛擬測驗服務中心之目標，則需建置 DAT (distance adaptive testing) 和 WWW 伺服器來完成使用者確認、選題、線上施測、試題分析、結果解釋及使用者界面等功能。

2. 題庫編輯

系統提供試題資料檔案上傳 (upload) 和線上編輯兩種試題編輯模式 (如圖四所示)，前者即讓測驗編製者依系統訂定之格式，先使用文字編輯器編製試題內容之文字檔，再上傳至系統；後者則提供線上之試題編輯介面，將試題資料傳至系統，由系統將試題資料儲存於暫存題庫中，並將試題資料檔案傳送給該領域專家及測驗專家審核，審核後之優良試題再由系統管理人員新增至題庫中，或與同類之自編測驗的題庫合併。

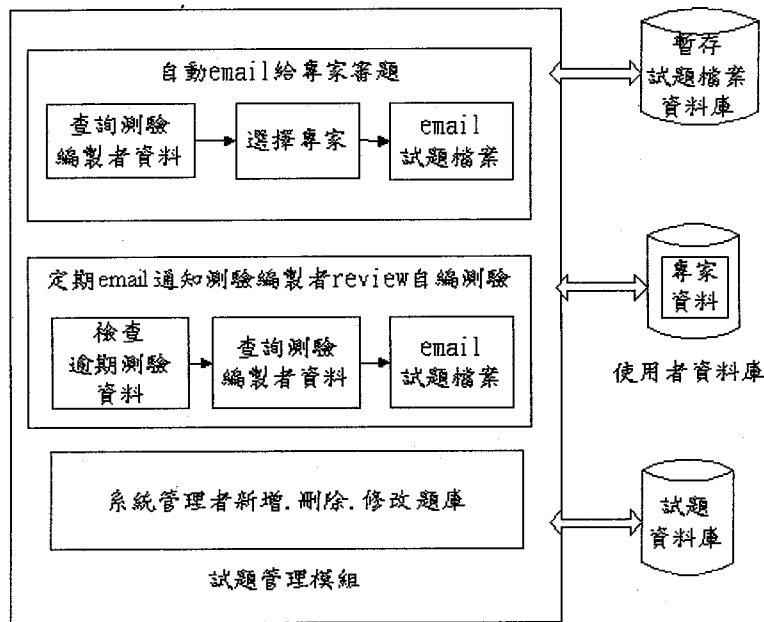


圖四 題庫編製流程

3.題庫管理

虛擬測驗服務中心需具備豐富、多元的題庫，以滿足使用者各項需求。為兼顧題庫品質的維持及避免系統管理者的沈重負擔，需由題庫管理模組協助系統管理者與測驗編製者相互配合，以進行題庫管理之工作(如圖五所示)。

題庫管理模組的主要工作包括自動將新增之試題資料，以 email 傳送給各專家進行試題之審查，定期通知各測驗編製者查看試題分析結果或回顧試題資料等。系統管理者則需對題庫進行新增、刪除、修改及查詢等工作。測驗編製者則透過系統提供之試題分析功能，了解自編測驗的品質指標(如古典測驗理論之試題難度、鑑別度等)及其使用情形，以改進命題技術，提昇測驗之品質，若發現試題品質下降成為不良試題，則通知系統管理人員刪除之。



圖五 題庫管理模組

4.題庫維護

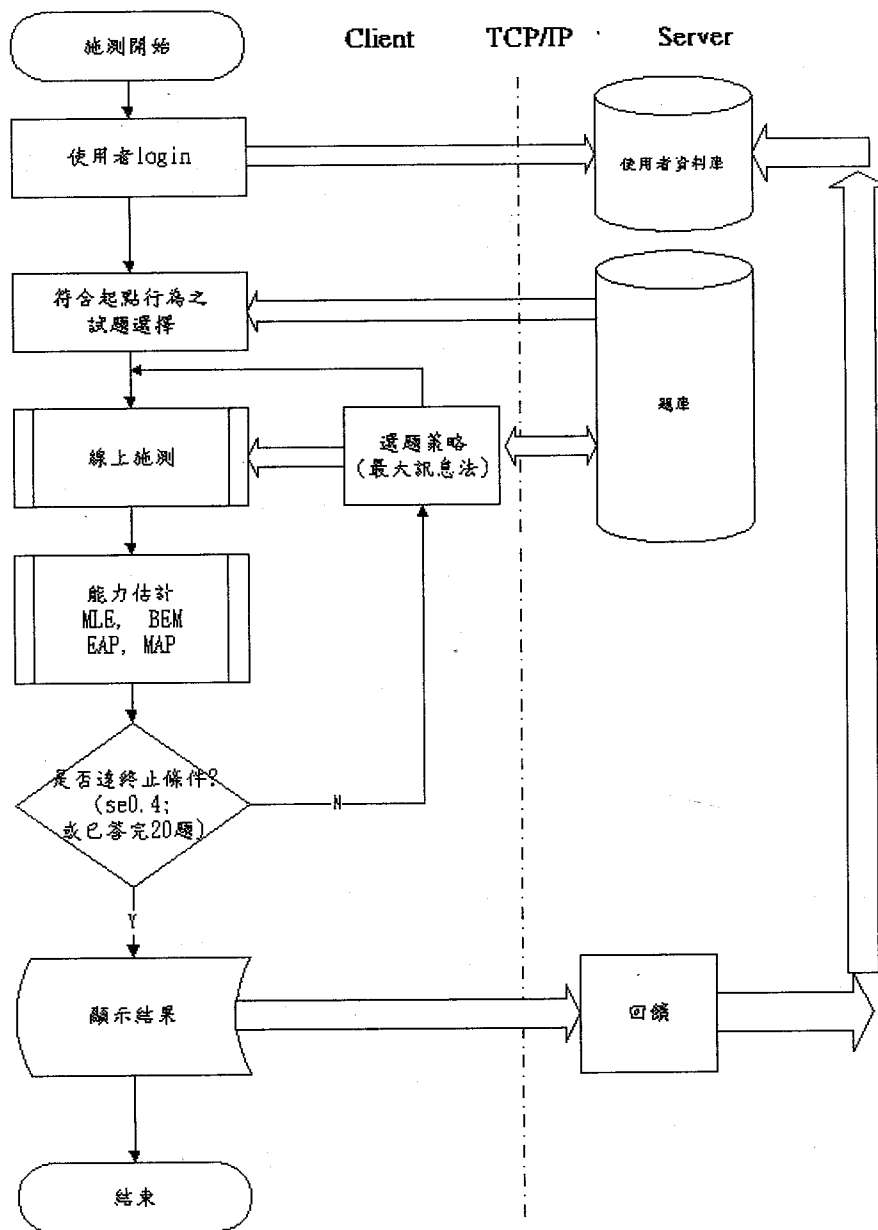
題庫經長期使用後，其品質會因課程調整、編製者不同、試題使用率差異等因素影響而改變 (Stocking, Martha, Lewis, & Charles, 1995)。為維持題庫之品質，本研究中加入試題曝光率控制機制之設計。另在非 IRT 題庫中之教師自編測驗，恐課程內容變更，使試題不符測驗目的，故採線上通知各測驗編製者定期回顧(review)測驗內容之方式，以維持其內容效度。

5.題庫安全

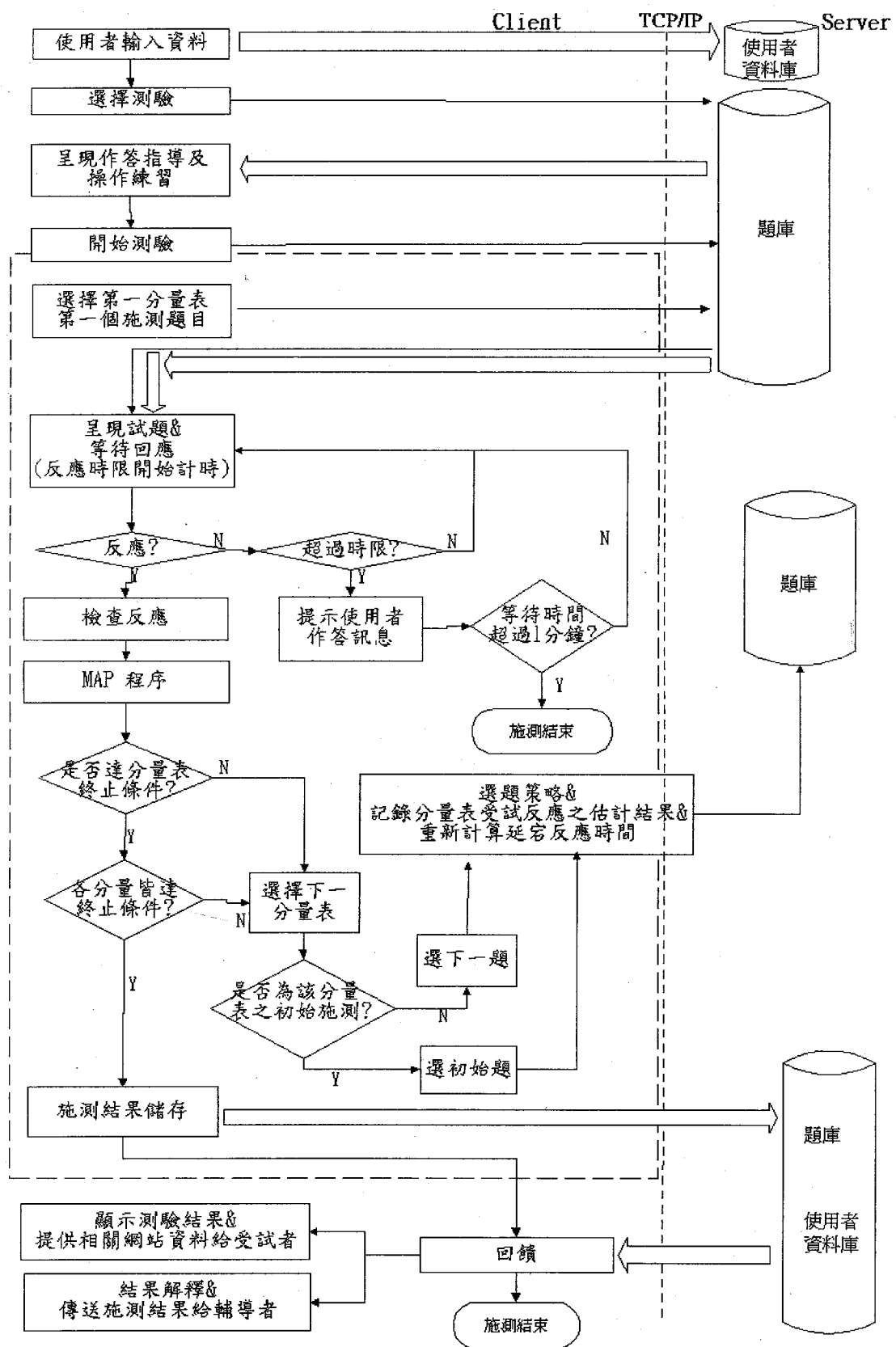
為防止題庫外洩問題或遭有心人士竊取，因此重視編製權限設定和題庫安全管理等功能之設計。在編輯權限設定方面，為防止試題編修管理產生混亂無序之現象，僅系統管理人員對全題庫具有管理權限。測驗編製者只能查詢自編測驗，對他人編製試題無查詢權限。系統則經由選題策略和試題曝光率控制機制，來維護 IRT 題庫之安全。

6.線上測驗

線上施測所使用的題庫包含傳統施測和適性施測兩類。前者為標準化測驗，採用古典測驗之逐題施測方式；後者又包含之最大表現及典型表現兩類測驗，兩者之實施程序皆為至題庫中選取題目，將測驗下載至使用者的 browser，進行線上測驗。而其反應組型、做答題數、受測時數等訊息將會經計算之後，連同估計結果一併傳回 server 端，存於使用者資料庫。但因兩者之施測目的和估計模式不同，故將兩者之實施程序差異部分標示於圖六及圖七中。



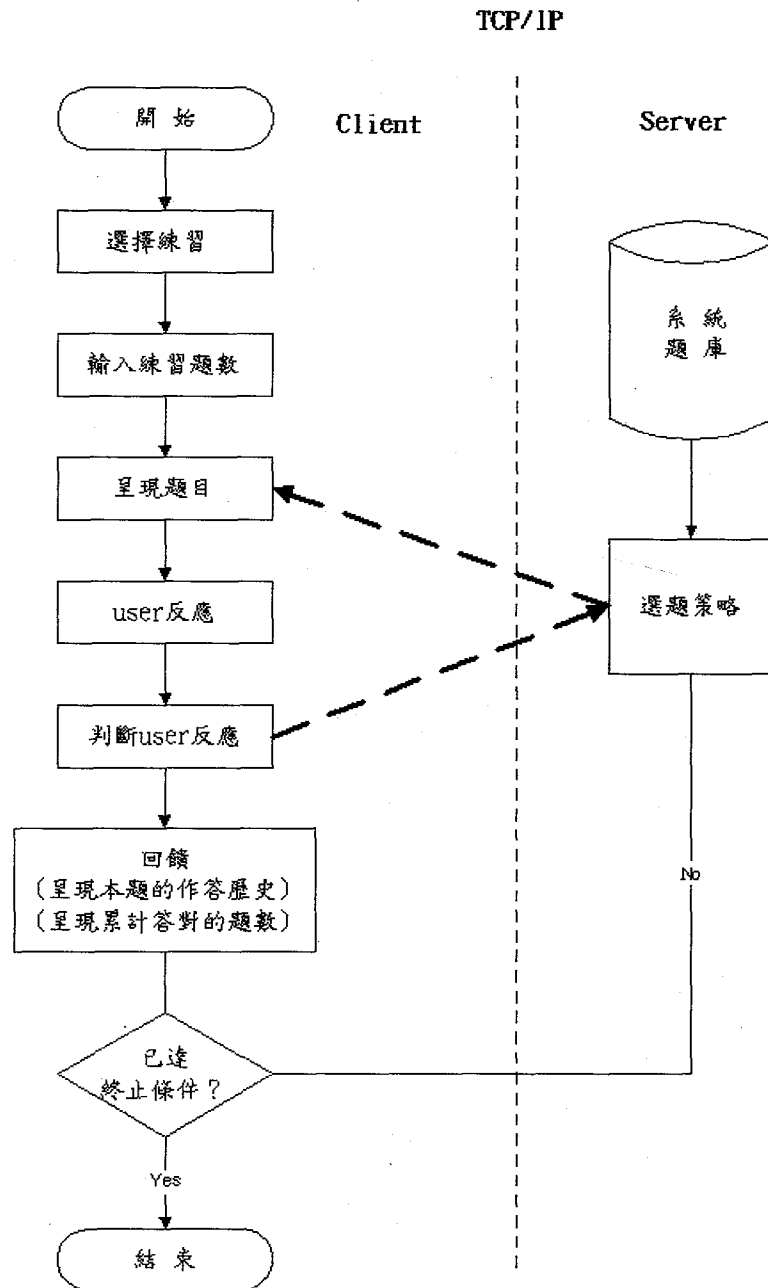
圖六 線上施測實施流程--最大表現測驗(適性測驗)



圖七 線上施測實施流程 -- 典型表現測驗 (適性測驗)

7. 線上練習

線上練習主要由非標準化題目構成，亦即提供無需常模的教師自行編製之練習題，實施模式採VIPQ策略，以收練習效果，其流程如圖八所示。



圖八 線上練習實施流程－教師自編測驗

8. 試題分析

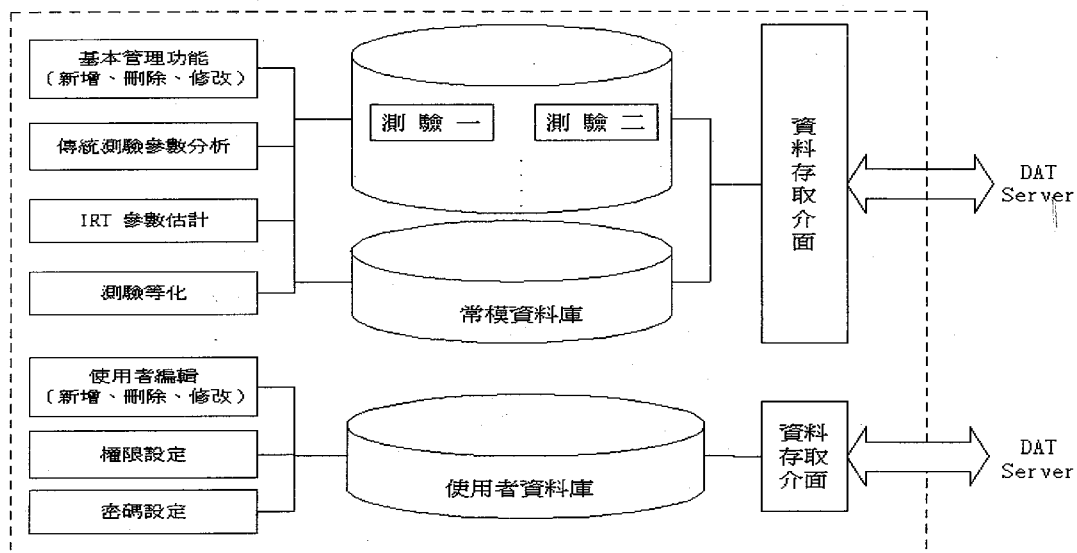
測驗編製者，可利用系統進行古典測驗或適性測驗之試題分析，其進行方式包括：(1)外部受試反應型組檔案輸入，(2)運用系統於線上收集之受試反應組型進行分析。

9.使用者資料庫之建立

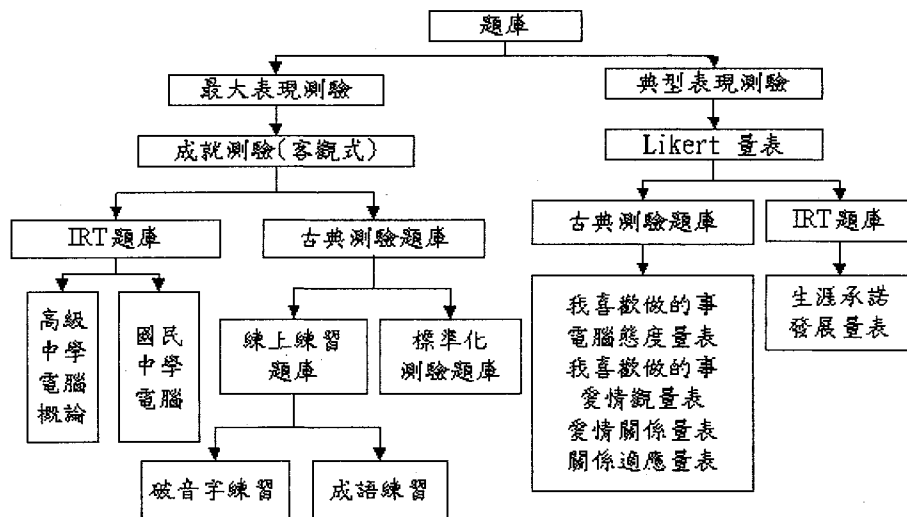
為儲存使用者之各項資料，便於收集線上施測反應，提供進行試題分析所需，也可做為日後系統功能之擴充，增加使用者查詢施測結果。其內容包括編號、使用過之測驗或練習編號、反應資料、email 信箱等資料，另典型表現測驗所需之受試者基本資料，各測驗不盡相同，所以在題庫建立時，才動態建立使用者資料庫之結構。

四、系統開發

目前系統雛型已建置完成(如圖九所示)，題庫之測驗內容包含六個典型表現測驗(五個古典測驗題庫，一個 IRT 題庫)、二個最大表現測驗(皆為 IRT 題庫)以及二個線上練習(教師自編測驗)，測驗名稱詳見圖十，而系統之試題編製、線上施測、線上練習等功能，限於篇幅，請參閱本系統之網址 <http://140.122.77.158/catsystem/>。



圖九 系統開發雛型



圖十 題庫內容

肆、結果與討論

虛擬測驗服務中心之研發工作，至目前已具雛型，各類題庫範例也建立完成，並持續建立各類題庫。適性測驗實施效率經模擬實驗的評估，結果顯示：最大表現測驗最少經 6 至 10 題之施測，即可正確計算受試者之能力估計值，最多則需 39 題(總數為 150 題)；典型表現測驗最少經 3 題之施測，可正確計算受試者之態度估計值，最多則需 28 題(總數為 46 題)。

目前對系統進行功能測試，試題編輯、線上施測、線上練習、試題分析等功能皆可正常運作。進行系統穩定性之測試，發現線上同時使用達 50 人時，系統仍可正常運作，而在網路傳輸非尖峰時段，選題呈現之反應時間約為 1 至 2 秒。由本研究之結果，可驗證虛擬測驗中心的建立是可行的，遠距適性測驗則可精確反應受試者之潛在能力或態度值，且能有效減少施測驗題數。

伍、結論與建議

目前國內尚無其開發類似虛擬測驗服務中心以提供測驗服務為主之系統，筆者們期望藉由虛擬測驗服務中心的建立，將來能有更多教師和測驗編製者的加入，或透過網路命題競賽活動舉辦，整合教學和測驗專業知能，提高試題編製的品質，建立更豐富的測驗資源網，充分發揮測驗的功能，並開創一個新的遠距測驗實施模式。

參考文獻

- 何榮桂(民 84)。練習式 CAI 選題策略之比較研究。師大學報。Vol.40 pp.77-110。
- Baker, F. B. (1992). *Item response theory: parameter estimation techniques*. NY: Marcel Dekker.
- Ho, R. G. (1989). Computerized adaptive testing. *Psychological Testing*, Vol. XXXVI, 117-130.
- Hsu, T. C. , & Sadock, S. F. (1985). *Computer-assisted test construction: A state of art*(TME Report 88). Eric on Tests, Measurement, and Evaluation. Princeton. NJ: Educational Testing Service.
- Millman, J. & Arter, J. A. (1984). Issue in banking. *Journal of Educational Measurement*, 21(4),315-330.
- Ramez, E. & Shamkant, B. N. (1994). *Fundamentals of database systems*(2rd). The Benjamin/Cummings Publishing Company, Inc.
- Stocking, Martha L., Lewis & Charles. (1995) *A new method of controlling item exposure in computerized adaptive testing. (Report.No95-5)*. Princeton, NJ: Educational Testing Service.