一、 緒論

1.1 研究動機

測驗在人類社會生活中，佔有十分重要的地位。舉凡國家之文官甄選、證照頒給、升學制度、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(何榮桂等, 民 86[1])。考試作為衡量人的能力的重要手段，在現代生活中的地位進一步提高，並深入到社會的各個層面。各式各樣的學歷考試、資格證書考試層出不窮。同時，隨著電腦、通訊、網路技術的不斷發展，考試的方式和媒介也發生革命性的變化。從傳統的紙筆考試到電腦輔助考試，到最新的植基於網站(web-based)的考試。就留學考試 TOFEL 和 GRE 而言，實為整個演變代表實例。如何運用最新技術，並客觀、準確地評估個體的知識和能力水準，已成為研究的熱門問題(申瑞民、曾華軍[2])。

「測驗與評量」一直是教學過程中不可或缺的一環，它們可以反應出學生在學習過程中了解知識的程度，以提供教學者掌握學生的學習狀況。傳統的測驗型態，一直是以文字、圖形的靜態紙筆測驗為主。隨著資訊科技的快速發展，對於傳統教學活動產生不小的衝擊，「資訊科技融入教學」活動目前正如火如荼的展開。因此，具有教學成果回饋與評鑑功能的電腦輔助測驗也逐漸地在發展(孫光天、陳新豐、吳鐵雄，民 87 [3])。九年一貫課程中英語科強調培養學生聽、說、讀、寫能力（proficiency）。客觀的英語科學習成就評量需要新的測驗方式。英語科的電腦化測驗（computerized testing, CT）受惠於電腦硬體效能不斷提昇及語音辨識軟體的普及，並且能在測驗過程中與學習者產生高度互動性，將成為英語教育上的一大利器，例如：美國著名的 GRE 測驗及台灣地區的托福考試分別在西元 1999 年及西元 2000 年全面廢止舊行的紙筆測驗方式，改用電腦來進行考試，可見藉由電腦來進行測驗已是當前的趨勢。

電腦化測驗所依據的理論可分為兩類，在過去多以古典測驗理論為主，但由於古典測驗理論是對所有的人施測同一組題目，這對有些類型的測驗來說不是很恰當，例如：學習能力測驗，由於每個人的學習能力並不相同，若不考慮個別差異而給予相同的題目，對能力水準較高的人來說，大多數題目可能顯得太簡單了，不但作答過程過於乏味，測驗結果也無法正確反映其能力；反之，對能力水準較低的人來說，則大多數題目可能太難了，不僅作答時容易感到挫折與焦慮，大多數的題目也可能僅以猜測的方式來填答，同樣無法正確測量到其能力，所以試題難度及鑑別度亦會因測驗樣本能力的不同而改變。另外，尚有測驗複本不易平行（兩次測驗上得到完全相同結果）、缺乏試題訊息…等缺點，因此在應用上較不適切。為解決古典測驗理論的缺點，試題反應理論（Item Response Theory, IRT）興起於 1950 年代以後，經歷數十年逐步發展，直到 1980 年開始的十年臻

於完善。

IRT 理論所發展之適性測驗，不僅適用於精熟式適性測驗，且適用於適性成就測驗或人格測驗，半世紀以來普遍受到重視與應用。IRT 理論具有三大特色：（1）在不同能力上，使用不同之估計標準誤，（2）題目難度與受試者能力都建立在同一量尺上，（3）參數之估計值具不變性。這三大特色使得適性測驗更精確與容易實施，讓接受不同題組之受試者能力也可直接相互比較（李茂能(民 89)[4]）。

適性測驗（Adaptive Testing）之發展可簡單依照施測方式分為四個階段：人工化、電腦化、智慧化、遠距化。試題反應理論（IRT）理論興起後，推動電腦化適性測驗（CAT）的應用與發展。1993 年美國國防部首先大規模採用三軍職業性向測驗（CAT-ASVAB）於軍隊人事之篩選與安置上，對 CAT 測驗之推廣與應用於大規模測驗上具有示範作用。至此國內外適性測驗之理論基礎雖已漸趨具體明確，但施測、計分方法仍甚為繁瑣及費時費力。直到 90 年代價格便宜的個人電腦普及之後，電腦化適性測驗才具體可行易於實際應用。電腦化適性測驗逐漸普遍為國外許多證照考試、商業機構、測驗公司、公私立學校、與軍方所應用（李茂能(民 89)[4]）。

電腦化適性測驗的最大優點就在於它能以最少的試題準確評量出受測者的能力，因為所施測的每道題目都反映了它對受測者能力的最新估計，所以對受測者來說，測驗的題目既不致太難也不會太簡單，而使得評量結果的準確性得以大幅提升。因此使用適性測驗，不但可因施測題數減少而降低施測的時間成本，同時測驗結果的準確性也可以獲得提升。但是以試題反應理論為依據的電腦化適性測驗，所使用的樣本通常需要 200 至 1000 名受試者參與，以校準其題庫中的試題參數。例如：日本的電腦化適性測驗實例〝CASEC〞（Computerized Assessment System for English Communication）。CASEC 所使用的題庫（item bank）包含三參數（CASEC 測驗第一～第三部分：字彙、片語、聽力）與雙參數（CASEC 測驗第四部分：聽寫填空）4000 題試題。這些 IRT 試題參數的校準是經歷 10 次以上的取樣測驗（prestest）才完成。每次取樣測驗是由六份試卷所構成，每份試卷包含 120 試題，六份試卷皆包含一部份共同的試題—共通題（common items），每份試卷參與的受試者（examinees）至少有 1000 人。取樣測驗結束後，利用共通題進行六份試卷的測驗等化（Test Equating）（N. Hayashi et al.，2004 [5] ）。CASEC 的取樣測驗耗費如此龐大的人力、物力，使得以實施傳統紙筆測驗形式取樣測驗的電腦化適性測驗，幾乎不可能按照 CASEC 模式套用於一般學校測驗情境。

有些遠距電腦適性測驗系統雖然結合了項目反應理論來提供線上適性測驗，但是為了建置符合試題反應理論模式的題庫，首先仍必須得到大量的受試者進行一般的紙筆測驗預試，以便蒐集大量的受試樣本，然後再利用文書編輯工具輸入所有樣本資料，最後再藉由IRT套裝軟體（例如：BILOG、MULTILOG、

MICROCAT…等等）去估算試題的IRT參數，缺點是整個過程仍然耗費大量的人力與物力。

如果能利用網際網路的特性，直接在線上陸續蒐集上線受試者的反應組型，系統管理者隨時監測樣本數目，待收集足夠樣本後，再啟動線上試題參數估算機制，或系統自動估算試題參數(何榮桂、賴信仁，民 86[6]、黃坤泉，民 90[7]、朱秦利，民 91[8])，順利完成試題參數線上校準。如此，便能解決電腦化適性測驗和某些線上適性測驗收集受試樣本的缺點。

電腦化適性測驗的主要特徵除了施測題目因人而異外，其測驗的長度也可因人而異，電腦化適性測驗的施測可以持續到對受試能力估計的準確性達到某一事先設定的標準為止。測驗的目的和性質不同，其終止標準可以有高有低，也可同時設定一個以上的終止標準。常見的終止標準是並用最大測驗長度（長度多在二十題到三十題之間），和固定標準誤--測量標準誤變動或後驗標準差(posterior standard deviation, PSD)小於預定的值。

一組受試反應組型若造成能力估計引擎在最大測驗長度時仍然無法正常收斂，基於測驗效率等因素考量，儘管能力估計未達預定精確度，系統必須提前終止適性測驗。因此電腦適性測驗應用最大測驗長度終止測驗所衍生的缺點是會造成能力估計誤差。換言之，對某些受試反應組型而言，測驗效率與能力估計誤差不可兼得。

所以，為了解決上述問題，變成本研究主要動機。

1.2 研究目的

本研究主要目的如下：

一、 建置一個可以線上收集取樣測試樣本之多媒體線上電腦適性測驗系統，除了透過網際網路累增（accumulate）樣本數之外，尚可利用系統之取樣試卷匯出與匯入功能，以檔案方式匯總測試樣本。

二、 測驗系統內建模擬適性測驗程式，模擬系統裝載不同能力估計引擎實際運作。此研究工具被用來比較四種能力估計引擎工具的能力估計效能。比較方法是四種能力估計引擎分別饋送多組特定受試反應組型，在不考慮測驗效率下，輸出個別受試反應組型的收斂能力估計值或題庫用盡時的能力估計值和測驗長度。並且進一步分析慢速收斂(slowly converge)或題庫用盡仍無法收斂(diverge)的暫時能力估計值變化(behavior of provisional ability)圖形。最終目的是提出解決某些特定受試反應組型測驗效率與能力估計誤差不可兼得的建議。

1.3 研究範圍

在 IRT 試題線上參數校準研究方面,本研究將引用孫光天教授研究成果—以類神經網路進行題目反應理論試題參數估算（孫光天、蔡志煌，民 89[9]、蔡志煌，民 89[10]），故試題參數估計不在研究範圍內。本研究將重點放在研究目的一。

在適性模擬測驗中，本研究將引用陳新豐，民 87[11]研究結果—160 題 IRT 試題參數，故題庫建置不在本研究範圍內。本研究將重點放在研究目的二。

1.4 研究工具

本研究計畫經智勝國際公司授權研究者針對其公司開發之『智勝鮮師教學網』現況之不足，配合研究目的進行系統延伸(enhancement)。因此，本研究所使用的主要研究工具--多媒體線上適性測驗系統的主體實為智勝國際公司的智勝鮮師教學網。它再經由研究者經由文獻探討國內外試題反應理論及電腦化適性測驗相關理論，以及配合研究目的，作為系統延伸的依據。經實際製作系統延伸部分，並經由系統測試後，完成本研究工具。