

# 線上多媒體適性測驗 取樣測驗之樣本收集與四種能力估計法模擬適性測驗比較

學生：毛家驥

指導教授：陳登吉 博士

國立交通大學理學院網路學習碩士在職專班

## 中文摘要

本研究將探討如何延伸植基於網站(web-based)之多媒體適性測驗系統，讓此系統擁有下列的特點：

一、線上收集適性取樣測驗樣本：透過適性取樣試卷檔案的匯出與匯入(export and import)功能，完成適性取樣測驗樣本之收集。當收集樣本數滿足估算試題參數條件時，便可啟動線上項目反應理論(Item Response Theory, IRT)試題參數估算功能。

二、四種能力估計引擎比較：系統內建 ASP 模擬適性測驗程式。主要程式元件中關於能力估計引擎(ability estimator engines)部分，則是分別利用四種不同能力估計法(即 OWEN、EAP、MLE、和 WLE 估計法)實作而成。系統依序將 50 組不同之特定受試反應組型饋送給能力估計引擎，進行電腦適性測驗能力估計模擬研究。

本研究結論如下：

- 一、 本研究所延伸之多媒體線上適性測驗(CAT)系統，能透過網際網路收集取樣測驗樣本，所新增之取樣試卷的匯出、匯入功能，便於在中小學校園環境實施線上取樣樣本收集工作。因此，避免傳統取樣紙筆測驗耗費花費人力、物力的缺點。
- 二、 貝氏能力估計法 OWEN 和 EAP 分別對一些特定受試反應組型，產生慢速收斂(slowly converge)或題庫用盡時能力估計無法收斂情況。
- 三、 最大概似估計法(MLE)和權值概似估計法(WLE)對於上述相同特定異常受試反應組型，其能力估計是正常收斂的。
- 四、 建議 CAT 系統運作發生上述反應組型時，採用最大概似估計法(MLE)重新估算。理由是：對於採用可變測驗長度(variable test length)CAT 系統而言，比較權值概似估計法(WLE)與最大概似估計法(MLE)的能力估計值偏差(bias)，以前者較大，這可從文獻探討與本研究之研究結果獲得證實。

四種能力估計法各有優缺點，例如：全對或全錯之反應組型適用貝氏能力估計法。因此，未來本系統將朝多能力估計引擎設計發展(例如：EAP+MLE引擎)，系統並具有自動偵測慢速收斂受試反應組型功能，自動切換適當能力估計引擎重新估算能力的機制，屆時測驗終止條件之一將可不再依賴最大測驗長度。在兼顧測驗效率下，以提高能力估計精確度為目標，屆時本線上多媒體適性測驗系統將變得更可靠。

# A Web-based Multimedia Computerized Adaptive Testing System — Online Pretest Samples Collecting and Simulating Comparison of Four Ability Estimation Methods

Student: Chia-Chi Mao

Advisor: Dr. Deng-Jyi Chen

Degree Program of E-Learning  
College of Science  
National Chiao Tung University

## 英文摘要

This study will explore how to enhance a web-based multimedia computerized adaptive testing (CAT) system. The enhancements of the multimedia CAT system include:

1. Online pretest samples collecting: Besides online pretest samples collecting, an offline pretest samples accumulation is accomplished by the files exporting and importing of CAT pretest samples. Whenever the count of some CAT pretest samples exceeds the system default, the online calibration of IRT items parameters might be manually enabled.
2. Simulating comparison of four IRT ability estimator engines: There is a system built-in functionality simulating real CAT sessions operating. It is implemented with ASP programs. The ability estimator engines are designed according to four IRT ability methods (OWEN · EAP · MLE and WLE ). The first step of the simulating study of four IRT ability estimators is to feed same specific response patterns to each of the four engines. Unlike an ordinary CAT system, maximum test length (e.g. 20 · 25 or 30, etc) is no longer served as one of test termination rules. For four IRT ability estimators, individual test length (could be the size of items bank) and estimated ability of each response pattern are of interest of this study.

The conclusion of this study is as follows:

1. After enhancing, the online multimedia CAT system pretest samples collecting feature avoids the drawback of huge examinees attendance and expensive cost caused by a conventional paper and pencil style pretest samples collecting task.
2. Bayesian ability methods, OWEN and EAP, would generate an outcome of slow convergence or running out of the items bank for some response patterns. On the contrary, maximum likelihood ability methods, MLE and WLE, would be typically convergent for these same response patterns.
3. To overcome the issue of contradiction between ability estimation accuracy and CAT efficiency caused by a specific response pattern for OWEN and EAP methods, it is suggested that a CAT system driven by a single ability estimation engine (e.g. EAP) is transformed into the multiple CAT ability estimation engines (e.g. EAP+MLE) scheme. Whenever the system detects a response pattern which could cause default engine to be slowly convergent, the mechanism of the multiple ability estimation engines would automatically switch to another engine to estimate ability again for the same response pattern.