

Computerized Adaptive Testing (CAT) 方式的 網路中文口語測驗的開發

村上公一、砂岡和子、劉松

日本早稻田大學

kimikazu@waseda.jp

關鍵字：CAT，IRT，電腦化適性測驗，試題反應理論，網路測驗

摘要：

電腦化適性測驗 (computerized adaptive testing, CAT) 是以試題反應理論 (item response theory, IRT) 為基礎的測試。根據考生的答題正誤，調整下一題的難易度，從試題庫中選出對學生最適合的試題。其優點在於：根據考生的答題正誤，調整下一題的難易度，所以考生不用解答過難或過易的試題，每試題對考生中文能力估測的訊息量很大，考生只解答相當於一般的試卷的 10%-20% 的試題及時間，就可以精確估計考生的中文水平。

早稻田大學 “Tutorial Chinese” 自 2003 年 4 月以來已嘗試過 4 次一般性的網路分班測驗，2005 年 4 月嘗試了第 1 次電腦化適性測驗 (CAT) 方式的網路中文口語測驗。

早大網路中文測驗共分為 3 個部分；

- (一) 詞彙聽力：共 10 題 (聽詞彙選日譯)，3 項選擇，用來測試考生的詞彙量
- (二) 完成說話：共 10 題 (聽一段話，然後補充後續部分，完成全文)，3 項選擇，用來測試考生的聽、說能力 (基礎表達能力)
- (三) 復讀：共 10 題，4 項自報選擇 (包括復讀全部、復讀一半、只復讀一兩個詞、一點也不能復讀)；用來測試考生的聽、說能力 (聲音化能力)

本文介紹我們開發的早大網路中文測驗的具體內容，並提供給中文教學者作為參考。

0・前言

早稻田大學從 2002 年開始推展 Tutorial Method 漢語會話訓練。目前有 213 名學生分 60 個小組參加訓練。推展互動型小班漢語教學時，為了確保同一班上的同學口語水平一致，以利分班，所以需要測驗學生的口語能力。我們自 2003 年 4 月

以來已嘗試過 4 次一般性的網絡分班測驗，2005 年 4 月嘗試了第 1 次電腦化適性測驗（CAT）方式的網路中文口語測驗。本文介紹我們開發的早大網路中文測驗的具體內容，並提供給中文教學者作為參考。

1・一般性的網絡分班測驗的概要

我們嘗試過 4 次一般性的網絡分班測驗；第 1 次在 2003 年 4 月及 7 月，第 2 次在 2003 年 12 月，第 3 次在 2004 年 4 月，第 4 次在 2004 年 7 月舉行。各次測驗的具體出題形式如下(表 1)。

	(1)詞彙拼音	(2)詞彙聽力	(3)回答提問	(4)完成說話	(5)復讀
第 1			共 5 題	共 5 題	共 10 題
第 2			共 5 題	共 5 題	共 10 題
第 3			共 5 題	共 5 題	共 10 題
第 4	共 10 題	共 10 題		共 10 題	共 10 題

表 1(網絡分班測驗出題形式)

已試過 5 種試題。(1)詞彙拼音；看詞彙選拼音，(2) 詞彙聽力；聽詞彙選日譯，(3)回答提問；聽一段話，然後回答提問，(4)完成說話；聽一段話，然後補充後續部分，完成全文，(5) 復讀；聽一段話，然後復讀，自報結果（包括復讀全部、復讀一半、只復讀一兩個詞、一點也不能復讀）。

第 3 次測驗在 2004 年 4 月舉行，108 名學生參加。同時讓學生參加 TECC(Test of Communicative Chinese)團體考試（4 月 24 日舉行）。TECC 是由(日本)中文交際協會組織實施，主要檢定考生的中文運用、交際能力的“中文交際能力測驗”。從 1996 年研制開發，到現在共計舉辦 16 次公開測試，另舉辦好多次個別團體測試。考試時間為 80 分鐘，聽力試 70 題、筆記試 70 題，共 140 題。108 名中 78 名參加 TECC 團體考試。我們分析 2004 年 4 月份舉行的兩種測試的學生成績的相關性，確認分班測驗的可靠性。表 2、表 3 是第 3 次分班測驗和 TECC 團體測試的基本統計表和相關系數表。

	(3) (25)	(4) (25)	(5) (50)	總分 (100)	聽力 (500)	筆記 (500)	總分 (1000)
平均	9.598	10.920	28.080	48.598	227.310	168.092	395.402
標準偏差	5.868	4.733	11.182	18.286	69.046	57.744	121.477
最低	0	0	9	11	92	76	193
最高	0	25	50	90	436	396	832

表 2(基本統計表)

	(3)	(4)	(5)	總分	聽力	筆記	總分
(3)	1.000						
(4)	0.411	1.000					
(5)	0.609	0.470	1.000				
總分	0.800	0.678	0.929	1.000			
聽力	0.561	0.513	0.743	0.767	1.000		
筆記	0.456	0.443	0.645	0.656	0.835	1.000	
總分	0.536	0.502	0.729	0.748	0.965	0.950	1.000

表 3 (相關系數表)

第 3 次網絡分班測驗總分與 TECC 團體測試的總分的相關系數(R)達 0.748，屬於高度相關。與聽力試得分的相關系數(R=0.767)比與筆記試得分的(R=0.656)高 0.111，表示網絡分班測驗的總分主要反映以聽力為主的中文交際能力。

網絡分班測驗的三種出題形式中與總分的相關性最大的是 (5) “復讀”，其次是(3) “回答提問”，(4) “完成說話” 最小。“復讀” 與 TECC 總分、TECC 聽力的相關性也相當高，可以認為是最合適的出題形式。因為“回答提問”和“完成說話”的出題數只有 5 題，所以與 TECC 總分、TECC 聽力的相關性都不那麼高。

按分班測驗的結果分班後開始 Tutorial 課。第一週的 Tutorial 以自我介紹為中心，上課時 Tutor 確認各各學生的實際的中文交際能力的小班上的次序。表 4 是 Tutor 課上確認的學生次序與按照各種測驗的分數推測的次序的比較。

“0” 表示 Tutor 確認的次序與按照各種測驗的分數推測的次序一樣，“+”、“-” 數字表示兩種次序之間的差異。按照 TECC 聽力推測的次序和網絡分班測驗推測的次序與 Tutor 確認的次序比較接近，但不少學生的次序有±2 以上的差異。

表 5 是 Tutor 課上確認的學生次序

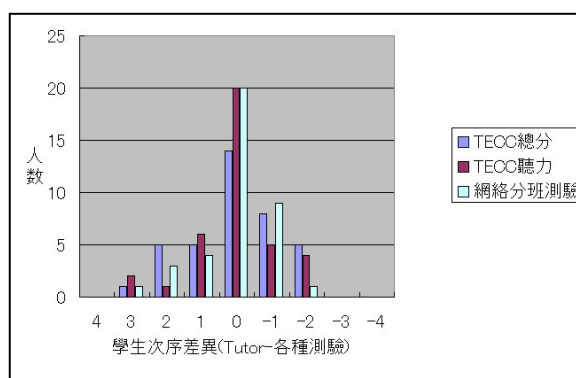


表 4 (學生次序差異 1)

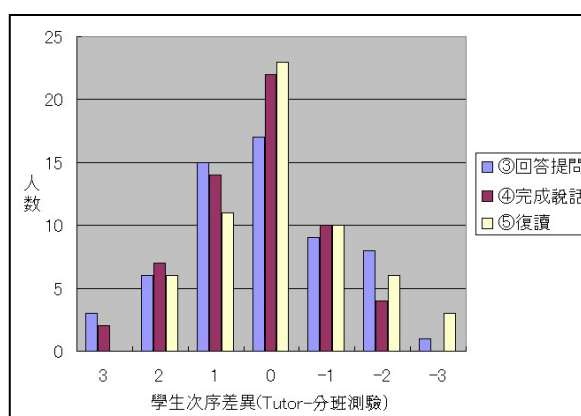


表 5 (學生次序差異 2)

與按照網絡分班測驗的各種出題形式的分數來推測的次序的比較。按照完成說話、復讀來推測的次序與 Tutor 確認的次序比較接近。回答提問差異比較大。

通過這些比較分析，我們了解“回答提問” 5 題、“完成說話” 5 題、“復讀” 10 題的一般性的網絡分班測驗已有和 TECC 差不多一樣的水平，但還不到我們所需要的精度。

為了提高測驗精度，我們自 2004 年 4 月開始與早稻田大學 International 公司、教育測定研究所合作研制開發電腦化適性測驗（CAT）方式的新的網路測驗。

2·電腦化適性測驗（computerized adaptive testing,CAT）

電腦化適性測驗（computerized adaptive testing,CAT）是以試題反應理論(item response theory, IRT)為基礎的電腦測試。

試題反應理論與古典測驗理論的最大差異在於；後者依據各各試題的實分數，前者依據各各試題的參數。先實施大量預試，算出各各試題的難易度、鑑別度、猜測度等參數，建立題庫，然後自題庫中挑選試題以組成測驗。考試完後，處理預先算出好的參數推測考生的能力。因為題庫裡的試題的參數都算出在同一個量尺上，一旦建立試庫我們就可以編製好幾種“同一個量尺、不一樣內容”的測驗。近幾年研制開發的外語測試一般都採用試題反應理論。上面所提的日本 TECC 也採用這些方式。

適性測驗(Adaptive testing)是利用依據試題反應理論建立的題庫組成測驗的一種方式。根據考生先前的答題情況，自題庫中挑選最能評量出考生能力的下一題進行施測。比方說如果考生上一題答對了，下一題就選難一點的試題，答錯了就選簡單一點的試題，每次都從試題庫中選出對學生最適合的試題。

傳統方式的測驗一般讓所有考生考同一張試卷，不考慮考生的個別差異，難題、易題都放在裡面，考生常碰到過難、過易的試題。這些試題對估測考生能力沒有積極作用，只是消耗考生的時間和精神而已。適性測驗的每道試題則都根據考生的答題正誤，調整下一題的難易度，所以考生不碰到過難、過易的試題，每試題對考生中文能力估測的訊息量很大，可以節省試題量及時間，而提高測驗的準確性。適性測驗由於選題、計分數需要大量的運算，因此適於電腦實施，這叫做“電腦化適性測驗（computerized adaptive testing,CAT）”。

下面介紹我們研制開發的電腦化適性測驗（CAT）方式網路測驗的建設流程及具體內容。建設流程如下。

- （一）測驗總設計(確定測驗的目的、形式、出題數、精度等)
- （二）試題命題(為了預試)
- （三）實施預試(利用一般性的網絡分班測驗系統)
- （四）建立試題庫(算出各各試題的難易度、鑑別度、猜測度等參數)
- （五）設計測驗系統

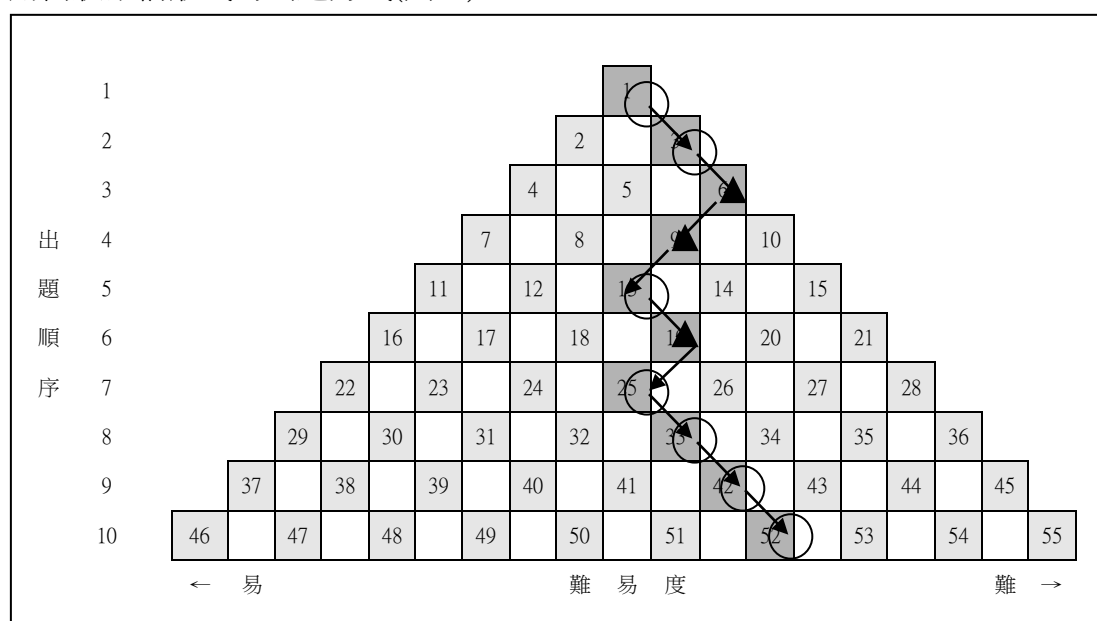
測驗的主要目的是精確估計以表達能力為主的中文口語水平，作為分班測驗、期末測驗之用。我們在第 1 到第 3 次一般性測驗已試過“回答提問”“完成說話”“復讀”的 3 種出題形式，第 4 次測驗再加上“詞彙拼音”及“詞彙聽力”兩種詞彙問題，我們採用比較反映考生的口語能力的“詞彙聽力”“完成說話”“復讀”的 3 種出題形式。

(一) 詞彙聽力：共 10 題（聽詞彙選日譯），3 項選擇，用來測試考生的詞彙量

(二) 完成說話：共 10 題（聽一段話，然後補充後續部分，完成全文），3 項選擇，用來測試考生的聽、說能力（基礎表達能力）

(三) 復讀：共 10 題，4 項自報選擇（包括復讀全部、復讀一半、只復讀一兩個詞、一點也不能復讀）；用來測試考生的聽、說能力（聲音化能力）

詞彙聽力及完成說話依據電腦化適性測驗方式，復讀採用傳統形式（但因隨機存取而出題，所以每次試題不一樣）。各 10 題，共 30 題。詞彙聽力及完成說話採用樹狀結構形式的出題方式(圖一)。



圖一(樹狀結構出題形式)

圖上○表示答對，▲表示答錯。考生答對了第 1 題（試題 1），就出題比試題 1 難一點的試題 3。答對了第 2 題（試題 3），就出題比試題 3 難一點的試題 6。答錯了第 3 題（試題 6），就出題比試題 6 簡單一點的試題 9。每道題都根據考生的上一題的答題正誤，調整難易度，所以考生不用解答過難或過易的試題。圖上的學生分為 19 層難易度的試題中只回答 4 層試題（〔1、13、25〕〔3、9、19、33〕〔6、42〕〔52〕）而已。除了這 4 層以外的 15 層試題對這個考生來說過難或者過易的試題，對估測考生能力沒有用處。

我們在 2004 年秋期實施總試題數 270(詞彙聽力 120、完成說話 120、復讀 30)，

400 多名學生參加的預試，算出各各試題的參數，建立好試題庫，2005 年 4 月嘗試了第 1 次電腦化適性測驗（CAT）方式的網路中文口語測驗稱謂 “Test of Ability for Chinese Communication(TACC)” （試用版本）。

圖二（TACC 封面）

在 2005 年上學期開始的 4 月中旬，總共有 213 名學生考這個新網絡分班測驗，按測驗成績把 213 名學生分到 60 個小班。剛上完一週的小班課，還沒有學生因中文水平不適而需要換班。

表 4、表 5 是四月份進行的測驗的基本統計表及相關系數表。

	詞彙聽力 (40)	完成說話 (40)	復讀 (20)	總分 (100)
平均	22.023	22.032	12.361	56.418
標準偏差	6.122	5.319	4.511	14.012
最低	7	10	1	26
最高	40	40	20	92

表 4（基本統計表）

	詞彙聽力	完成說話	復讀	總分
詞彙聽力	1			
完成說話	0.682	1		
復讀	0.638	0.637	1	
總分	0.901	0.883	0.843	1

表 5（相關系數表）

因為時間限制，我們還沒完全分析過這次測驗的詳細結果，也沒有進行過與 TECC 等其他測驗比較分析，但可以看得出這次測驗獲得了相當好的成果。

3・結語

我們自 2002 年開始推展 Tutorial Method 小人數漢語會話訓練以來，怎樣能進行最合適的分班，一直是重要的問題。這次我們開發的“Test of Ability for Chinese Communication(TACC)”是對這個問題的初步答案。只費 10 分鐘左右的測驗時間能做出相當好的分班結果。但 TACC 並不是完美的測驗，裡面有很多問題。比方說復讀題不是客觀分析考生的聲音評估的，目前採用自報選擇方式評估。這是一個暫時的措施。等到由於聲音分析技術的迅速發展，再等幾年一定能做出客觀評估。那時候我們也可以用新的技術評估復讀題。

附記：本稿發表得到自2002 到2004 年度日本文化科學省學術研究費(基盤研究B-2：研究題目“漢語Tutorial System 的國際開發” 課題號碼143101136：研究代表 砂岡和子)的撥款，是該研究成果的一部分。

參考書目

- 1・芝祐順(1991) “項目反応理論-基礎と応用-” (東京大学出版会)
- 2・村上公一、砂岡和子、劉松(2003) “中国語レベル分けのための Web 上での中国語能力診断テスト” (日本中国語学会第 53 回全国大会予稿集)
- 3・砂岡和子、村上公一 “以提升交際能力為目的之漢語教學以及水平測驗芻議-早稻田大學國際漢語遠距教學四年實踐見證”
- 4・村上公一、砂岡和子、劉松(2003) “発話能力を測定するための中国語 Web テストの開発” (日本中国語学会第 54 回全国大会予稿集)