出國報告（出國類別：參加國際會議）

參加 2009 年
美國研究生入學管理委員會(GMAC)
適性化測驗會議出國報告

服務機關：考選部
姓名職稱：蔡科長蕙仲、楊設計師淑如
派赴國家：美國
出國期間：98 年 6 月 1 日至 5 日
報告日期：98 年 9 月

參加 2009 年美國研究生入學管理委員會(GMAC)適性化測驗會議出國報告

摘　要

美國研究生入學管理委員會(Graduate Management Admission Council，GMAC)為全球商業學校的領導機構，其於 98 年 6 月 2 日至 98 年 6 月 3 日假美國明尼蘇達州之明尼亞波利市舉辦適性化測驗會議(2009 GMAC CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING)，該會議對於適性化施測結果與實務分析、推動適性化測驗所採用的程序方法、試題選擇之演算法、試題曝光率、多維度適性化測驗、題庫與題組發展、診斷測驗等研究皆有所著墨，並對於全球適性化測驗之研究與應用、以色列與西班牙國家對於適性化測驗發展之演進、美國政府支援適性化程式與專案發展之作法亦有所介紹。

透過參與此會議學習到適性化測驗最新理論與研究方向、國外政府機構支持適性化測驗之作法、各國推廣適性化測驗之現況，並和與會專家學者建立聯繫管道，於會議結束後擬撰出國報告，此報告針對會議部分議題作摘要式描述，並對以色列高等教育及美國研究生入學管理委員會(GMAC)發展適性化測驗經驗作詳細之報告，最後則針對定期參與國際會議、規劃與建置適性化測驗需考量之要項、後續之實地觀摩與考察提出心得與建議，期對於國家考試 e 化推動有所助益與貢獻。

關鍵字：適性化測驗、紙筆測驗、試題反應理論

目　　錄

表　目　錄

圖　目　錄

壹、前言

一、 背景緣起

考選部為推動 e 化電子政府，自民國 93 年起開始舉辦電腦化測驗，並擇定專門職業及技術人員特種考試航海人員（以下簡稱航海考試）考試採用電腦化測驗方式辦理考試，此項考試於北部考區國家考場電腦試場舉辦，一年辦理四次，考試試題除文字類型外，尚包含表格、彩色圖片等多樣化試題，且具備即測即評、亂題亂序等電腦化考試優點，充分結合考試需求與電腦特色，對於推動政府 e 化具有劃時代之意義。此外，考選部為逐步提升電腦化測驗實施效能並考量投資成本效益，自 96 年 7 月起擴大舉辦國家考試電腦化測驗，將專門職業及技術人員特種考試牙醫師、助產師、職能治療師、呼吸治療師、獸醫師考試等 5 類科（以下簡稱牙醫師等項考試）納入實施範圍，並於北部、中部、南部考區同時舉行，一年辦理兩次考試，以積極服務應考人。自 93 年舉辦電腦化測驗開始，截至 98 年 8 月止，已成功辦理了 22 次航海考試及 5 次牙醫師等項考試，成效斐然。

考選部為更積極推動政府 e 化作業，邁向全球先進之國家考試測驗機構，於考選部 98 年至 101 年中程施政計畫中明訂「精進考選技術提昇考試信度與效度：研發評量方法，擴大電腦化測驗考試範圍」之策略目標；另鑑於當代測驗理論中對於施測方式已漸漸興起一波電腦適性化測驗之國際潮流，其基於試題反應理論(Item Response Theory,IRT)所發展之施測方式與結果，在考試信度與效度上有其公正客觀之量化驗證方法，對於國家考試 e 化推動，極具參考意義與價值；爰此，為推動考選部為全球首屈一指之公正客觀政府測驗機構，並與世界潮流相結合，故派員參加國際適性化測驗研討會。

二、 目的

美國研究生入學管理委員會(Graduate Management Admission Council，GMAC)為全球商業學校的領導機構，其主管的美國商學研究生入學檢定考試(GMAT)被大多數美國高等教育機構所採用，用以評鑑申請入學者之能力等級，其對於美國商學研究生入學品質管理成效卓著。該委員會於 98 年 6 月 2 日至 98 年 6 月 3 日假美國明尼蘇達州之明尼亞波利市舉辦適性化測驗會議(2009 GMAC CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING)，該會議對於適性化施測結果與實務分析、推

動適性化測驗所採用的程序方法、試題選擇之演算法、試題曝光率、多維度適性化測驗、題庫與題組發展、診斷測驗等研究皆有所著墨，並對於全球適性化測驗之研究與應用、以色列與西班牙國家對於適性化測驗發展之演進、美國政府支援適性化程式與專案發展之作法亦有所介紹。鑑於美國研究生入學管理委員會(GMAC)舉辦之會議集結各國政府機構與專家學者分享適性化測驗發展經驗，並討論適性化測驗當下最新研究議題，爰此，考選部為了解適性化測驗最新理論與研究方向、學習國外政府機構支持適性化測驗之作法、參考各國推廣適性化測驗之現況，並蒐集全球有關適性化測驗之研究與應用，故派員參加是次適性化測驗會議。

三、 內容概述

本報告共分為前言、行程紀要、會議議題摘要、各國經驗分享、心得與建議及附錄六大章節，其中第一章節（前言）針對考選部派員參與 2009 GMAC CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING 之背景緣由、預期與會之目的加以描述；第二章節（行程紀要）則報告了赴美國參加會議之往返時間及各項行程；第三章節(會議議題摘要)則針對會議中部分研究做中文摘要式摘錄與報告，由摘錄之內容可以大略瞭解會議中有趣及實用之議題與內容，包含如何在適性化測驗中獲得較佳測驗結果、何謂最佳選擇之考試終止規則、美國軍職性向適性化測驗（Armed Services Vocational Aptitude Battery，ASVAB）之初步的介紹等；第四章節（各國經驗分享）則詳細報告以色列在辦理高等教育及美國研究生入學管理委員會(GMAC)在辦理美國商學研究生入學檢定考試(GMAT)時，其採用適性化施測方式之過程中、所考量之相關議題、程序及方法，並針對適性化測驗結果之公平客觀性有所報告；第五章節（心得與建議）則針對適性化測驗研習心得與建議事項提出報告，對於適性化測驗規劃時需考量之問題、建置時需研訂之演算法與方向及推動時所需面對之問題提出與會心得與建議，期望對於國家考試電腦化測驗有所貢獻。至於本報告最後一個章節-附錄，則檢附會議行程與議題原始資料（英文）、投稿論文摘要原始資訊（英文）、出國及與會行程之照片翦影供閱覽者參考。

貳、行程紀要

起程：由於大會舉辦地點位於美國明尼蘇達大學明尼亞波利校園的 Radisson 飯店內，爲了節省與會當日往返時程，並避免人生地不熟所造成的困擾，因此，有關住宿的安排，直接預定了大會舉辦會議的飯店，而參與會議起程，則於 6 月 1 日由台灣搭乘西北航空飛機，途經日本成田機場轉機，最後抵達明尼蘇達州明尼亞波利之聖保羅機場，至於機場至飯店之交通接駁，則透過當地相當方便的大眾運輸工具-輕軌車，由機場轉運至 MetroDome 巨蛋，再由 MetroDome 搭乘飯店接駁車抵達會議舉辦地點-Radisson 飯店。有關與會行程詳細時程表，詳如表 1。

參與會議：會議行程共有一天半，其中主辦單位提供兩天早餐及第一天午餐，會議中針對適性化測驗之研究議題或有實務面的分析（如：在適性化測驗中，當應考人不會作答時，究竟應該使用猜題的方式，或是應該跳過該題不作答），或有學術上演算法的研究（如針對多維度適性化測驗，有多篇研究針對多維度適性化測驗之相關演算法提出精進之改良與建議），甚至於第一天午餐時間之 poster session 中，亦有多國針對其適性化測驗之發展提出重點式介紹（如日本之 J-CAT、新加坡之 ESS 適性化測驗等），會議內容相當精彩豐富，不過，有關此次會議中之議題，鑑於此次與會目的主要係「藉由參考各國適性化測驗推動之策略、步驟、程序與方法，以習取各國豐富與經驗，進而加速國家考試 e 化之過程」，因次，在本報告章節中，將有一章節對此有詳細之描述。有關會議相關行程、議題與摘要資訊，詳如附錄一：會議議程、附錄二：會議投稿文章摘要。

城市建設觀摩：此次出國行程相當緊湊，對於美國城市建設之觀摩，僅可利用第一天及第二天會議行程結束後之剩餘時間在明尼蘇達大學校區及明尼亞波利市重點式參訪校園建築及城市建設，雖然時間緊迫，但也不可不謂是忙裡偷閒的一種休息方式。有關參加會議及城市建設觀摩的照片集錦，詳如附錄三。

不過在此要特別感謝台灣師範大學陳教授柏熹老師，在美國與會這一段時間，感謝陳教授充當城市建設導覽專家，並於會議上提供教育心理方面專業知識之協助。

表 1、參與會議行程

序號	日期	時間	說明	地點	備註
1.	6 月 1 日	06:00～07:20	起程（赴桃園國際機場）	台灣	
2.		07:20～09:35	起程（候機）	台灣	辦理出境及登機程序
3.		09:35～13:55	起程（轉機）	桃園國際機場～東京成田機場	西北航空(NW 22)
4.		13:55～16:25	起程（候機）	東京成田機場	辦理轉機程序
5.		16:25～13:20	起程（抵達美國）	東京成田機場～明尼亞波利聖保羅機場	西北航空(NW 20) 6/1 13:20 (美國時間)
6.		13:20～14:20	起程（入境）	明尼亞波利聖保羅機場	辦理入境程序
7.		14:20～15:00	起程（前往飯店-輕軌車）	明尼亞波利聖保羅機場～MetroDome	搭乘輕軌車
8.		15:00～15:30	起程（前往飯店-接駁車）	MetroDome～Radisson University Hotel	搭乘接駁車
9.		15:30～24:00	抵達飯店	Radisson University Hotel 615 Washington Ave Southeast Minneapolis	休息，準備參與第一天會議

序號	日期	時間	說明	地點	備註
10.	6 月 2 日	08:30～09:00	報到(參與會議)	參加會議 －第一天（上午） Radisson University Hotel (2F)	會議第一天
11.		09:00～10:15	適性化測驗真實面分析		
12.		10:15～10:30	中場休息		
13.		10:30～12:00	推動適性化測驗所採用的程序方法		
14.		12:00～13:00	午餐		
15.	6 月 2 日	12:30～14:00	全球對於適性化測驗之研究與應用	參加會議 －第一天（下午） Radisson University Hotel (2F)	POSTER SESSION
16.		13:00～13:40	西班牙國家對於適性化發展之演進		
17.		14:00～15:15	試題選擇之演算法 適性化測驗實務分析		CONCURRENT SESSION
18.		15:15～15:25	中場休息		
19.		15:25～17:30	美國政府支援適性化程式與專案發展之作法		
20.		17:30～24:00	城市建設觀摩、休息	明尼亞波利市、Radisson University Hotel	休息,準備參與第二天會議

序號	日期	時間	說明	地點	備註
21.	6 月 3 日	08:15～09:25	試題曝光率 多維度適性化測驗	參加會議 －第二天 Radisson University Hotel (2F)	CONCURRENT SESSION
22.		09:35～10:45	題庫與題組發展		
23.		10:45～11:00	中場休息		
24.		11:00～11:55	診斷測驗		
25.		11:55～12:30	總結與建議		
26.		12:30～24:00	城市建設觀摩、研讀會議資料	明尼亞波利市、Radisson University Hotel	休息
27.	6 月 4 日	07:30～11:20	校園建築觀摩	明尼蘇達大學校園 Minneapolis West Bank	準備返程
28.		11:20～12:00	返程（赴機場-接駁車）	Radisson University Hotel～MetroDome	搭乘接駁車
29.		12:00～12:30	返程（赴機場-輕軌車）	MetroDome～明尼亞波利聖保羅機場	搭乘輕軌車
30.		12:30～15:20	返程（候機）	明尼亞波利聖保羅機場	辦理出境及登機程序
31.		15:20～17:10	返程（轉機）	明尼亞波利～東京成田機場	西北航空 (NW 19) 抵東京途中跨國際換日線，需要加 1 日
32.	6 月 5 日	17:10～19:30	返程（候機）	東京成田機場	辦理轉機程序
33.		19:30～22:25	返程（抵達台灣）	東京成田機場～桃園國際機場-	
34.		22:25～23:00	返程（平安回家）	台灣	入境並返家

參、會議議題摘要

此次會議探討議程包含：適性化測驗真實面分析、推動適性化測驗所採用的程序方法、全球對於適性化測驗之研究與應用、西班牙與以色列國家對於適性化發展之演進、試題選擇之演算法、適性化測驗實務分析、美國政府支援適性化程式與專案發展之作法、試題曝光率、多維度適性化測驗、題庫與題組發展、診斷測驗等，相關研究與投稿文章約 50 篇，內容相當豐富，但限於報告完成期限與會議截稿期間，無法針對所有文章提供摘要式報告，故僅針對其中部分有趣或與考選部發展適性化測驗較爲相關之議題提出摘要式報告。

一、 考試長度的研究

適性化測驗之特色爲：考試一開始隨機自題庫中抽選起始試題，並依據應考人每次作答結果隨機自題庫中抽選出符合應考人能力之試題，直到符合考試終止條件時，即結束考試。會議中有一篇研究即針對施測終止之規則提出模擬研究（Termination Criteria in Computerized Adaptive Tests:Variable-Length CATs Are Not Biased），並提出最佳終止施測之方式，以供爲未來發展適性化測驗考試機構參考。

此研究鑑於以往研究多提出考試長度（時間）變動的適性化測驗（Variable-Length CATs）有較多的偏誤，故此研究以模擬方式檢視 4 個適性化測驗的終止規則（此四個終止規則分別爲：標準差 standard error，最小資訊量 minimum information，能力值改變 change in θ，固定考試時間 fixed length），另此研究並檢視由標準差與最小資訊量組合之終止規則，逐項分析不同適性化終止規則的估算能力、均方根誤（root mean square of error，RMSE），並且研究適性化測驗考試時間長度變動是否會真的造成偏誤(bias)。

此研究模擬過程中，能力值估算係採用最大近似法（Maximum likelihood）來估算能力值，且能力值以 0.5 逐量增加。至於研究結果說明如下：

1. 不論我們使用哪一種考試終止規則，考試時間較長的適性化測驗對於應考人的能力估算較正確，不過若題庫試題數量到達一定數量時，應考人能力估算之正確度與考試時間則不會有絕對相關，通常建議適性化測驗應該至少有 15-20 個試題來確保能力衡量的穩定度。
2. 若題庫包含足量的費雪訊息（Fisher Information），使其標準差非常小

時，則標準差終止規則將優於其他終止規則，且當管理相對少量的試題數量時，標準差終止規則也是一個相當有效率的方式。

3. 相較於固定考試長度的的終止規則，能力值改變（change in θ）這個終止規則表現較差。
4. 若題庫擁有高試題訊息函數，即使題庫試題數很少，混合終止規則（如標準差與最小資訊量之混合雙重終止規則）表現的最好。

此外，以往研究結果提出：「相較於考試時間固定的適性化測驗，考試長度（時間）變動的適性化測驗會產生較多的偏誤」，此研究指出該結論乃基於一些特殊研究設計下所產生出來之結果。最後，此研究發現相較於固定考試長度（時間）的終止規則，標準差的終止規則在低能力組的真實能力估算上有較佳的表現。

二、 猜題或跳答

本會議有一篇非常有趣的研究，其研究主題係探討在適性化測驗中，當應考人不會作答時，究竟應該是使用猜題的方式，還是應該跳過該題不作答，才能在適性化測驗中拿得較好的成績（Guess What? Score Differences With Rapid Replies Versus Omissions On A Computerized Adaptive Test），此研究背景乃基於適性化測驗的假設：「應考人僅依據他們所知的知識與技能來作答，作答之結果可適當地預估應考人之能力值」，研究資料係使用美國商學研究生入學檢定考試（GMAT）實際的適性化測驗資料來分析比較猜題與不作答在 GMAT 語言能力及數學能力測驗上之差異，期望提供給應考人一個得分作答的指引。

在尚未研讀此研究之建議前，應對 GMAT 考試有一個基本認識，GMAT 包含三大部分：(1)分析寫作評估（Analytical Writing Assessment，AWA）(2)數理能力評估（Quantitative section）(3)語言能力評估（Verbal section），其中數理能力及語言能力評估採用適性化測驗，若應考人答錯答案，將會倒扣分數。

此研究實驗中對於猜題的定義爲：「在考試結束前快速連續的猜測行爲」，至於究竟「考試結束前」及「猜測行爲」如何定義呢？此研究針對「猜測行爲」定義之臨界值爲：語言能力測驗上的猜題行爲須少於 10 秒鐘，數理能力測驗上的猜題行爲須少於 7 秒鐘，舉例來說，若有一位應考人在數理能力測驗考試結束前之最後四題

作答時間爲 4 秒鐘、12 秒鐘、3 秒鐘、5 秒鐘，則這個應考人的作答行爲僅認定爲最後兩題是猜題行爲（4<7、12>7、3<7、5<7，將試題由後往前觀察，僅兩個試題連續小於臨界值)。此研究之樣本僅擷取在考試結束前連續作答時間皆小於臨界值之作答資料，且其猜題題數介於 1 至 5 題（包含 1 題及 5 題）之資料。

經過實際資料的研究分析，基本上猜題對於應考人而言會是比較好的作答策略；不過在實證結果上，卻也必須依據試題數多寡、考試類別及應考人能力來做不同之作答策略，研究結果說明如下：

1. 大體上而言，猜題對於應考人而言會是比較好的作答策略
2. 在 GMAT 語言能力測驗上，猜題與不要作答的成績結果差異很小
3. 在數理能力測驗上，當試題數量變多時，猜題是較好的策略；
4. 若針對不同能力應考人來分析作答策略時，對於程度低的應考人而言，不論是在語言或是數理能力測驗上，不要作答會是比較好的策略；對於程度佳的應考人而言，在題數不多的數理能力測驗上，依據專業知識猜題是最好的策略。

最後，此研究的作者建議未來的研究不要僅針對考試結束前快速連續的猜測行爲作研究，可以著墨於隨機猜題或是整個考試過程中的猜題。此篇文章非常的有趣，且對於未來有興趣參與 GMAT 考試之應考人，可以在作答策略上做一番規劃。

三、 決策理論適性測驗

會議中上有一篇研究不以試題反應理論(IRT)爲適性化測驗之基礎理論，而使用決策理論來發展適性化測驗（Adaptive Testing Using Decision Theory），此研究提及在傳統教科書中，有些學者認爲測驗的目的主要是要達到分類決策，且在現在的工作環境中，有很多決策是二元的（如是否要僱用某人、受測者是否精熟於某項特殊技能等）。此外，還有一些測驗之主要目的係將受測者區別若干類(即目錄式分類，不同於連續性的結果。如將學生能力分類爲基礎能力、專業能力及進階能力３類，則測驗之結果係要評估有哪些比例的學生具備基礎能力、有哪些學生具備專業能力、有哪些學生具備進階能力等分類)。此研究提出試題反應理論(IRT)模式早已被應用在

分類式的決策上，惟其方法需大費周章地計算每個受測者能力值，然後再依據切割點，將每位受測者分類；此外，試題反應理論(IRT)並不永遠適用於實際狀況，因爲試題反應理論(IRT)相當地複雜、須符合多個嚴格的假設、需要大量校樣後的樣本，而當測驗目的僅是簡單分類決策時，試題反應理論(IRT)並不是很有效率。

因此，此研究依據衡量決策理論(measurement decision theory)提出一個適性測驗的模式，並使用模擬的試題反應資料，比較其與試題反應理論(IRT)模式在分類上的正確性。

此研究檢驗三種決策理論適性選題模式，此三種模式分別爲(1)傳統決策理論循序測試法(期望最小成本) (traditional decision theory sequential testing approach，expected minimum cost)、(2)資訊獲取(information gain)法及(3)最大鑑別力(maximum discrimination)法。研究結果顯示最小成本法顯著地比試題反應理論(IRT)好；而以資訊理論(Information theory)及熵(entropy)爲基礎的資訊獲取則與最小成本法不相上下；最大鑑別力雖然比上述兩者差，但是仍然比試題反應理論(IRT)好。此外，此研究發現若以 SPRT 爲測驗終止條件，90%模擬資料(模擬全國教育發展評估 National Assessment of Educational Progress (NAEP)之測驗資料)中有 86%可以被正確分類。因此，此研究提出來的決策理論模式是一個功能強大且可被廣泛應用的模式，他的好處是可以做分類、僅需要少量試題、非常容易建置、僅需少量前測、可以拿來做診斷測試、可以應用在多重技能的分類、可以讓一般使用者輕易地分析資料模式。

四、 不同施測方式之同等性

會議中有篇論文針對不同施測方式之同等性提出研究報告（Assessing the Equivalence of Internet-Based vs. Paper-and-Pencil Psychometric Tests），此研究鑑於使用網際網路作爲施測工具有逐漸上升之趨勢，其主因係因網路提供便利性與效率，然而適性化測驗與傳統紙筆測驗有許多差異性，以施測方式來看，適性化測驗運用資訊科技表現不同的試題類型，作答方式也不同，且閱讀能力測驗也可以運用資訊科技顯示圖文並茂的試題，並控管應考時間；然而，適性化測驗卻需要考量電源供應、非標準化電腦規格、伺服器效能、網路伺服器流量、試題被竊等問題，鑑於紙筆測驗與電腦適性化測驗有上述之差異性，因此此研究針對不同施測方式之同等性提出

研究報告。

此研究之資料來源係依據以色列高等教育機構招生時舉辦的教育心理測驗(Psychometric Test)結果做資料分析與驗證，教育心理測驗(Psychometric Test)在以色列是一個高度利害關係的考試，這個考試包含語言(60 題)、數理(60 題)與英語測驗(54 題)，所有題目是單選題，以往這個考試是紙筆測驗，現在則爲兩種施測方式（紙筆測驗與適性化測驗）同時進行，而未來則擬以適性化測驗爲主，爲了增加現今與未來推動適性化測驗之說服力，故此研究針對兩種考試方式的同等性(不論以何種施測方式，施測結果皆相同)進行研究，期望能驗證適性化測驗具有相同之測驗能力。

此研究以報名註冊 2008 年 10 月教育心理測驗考試的 381 位應考人來做研究分析，其中自 381 位應考人中隨機挑選 192 個應考人以紙筆測驗施測，189 個應考人以網際網路來模擬施測，此外，這兩組各選出 185 人，在此實驗的一個月後參加真正的教育心理測驗。經過資料分析，此研究發現以下的結論：

1. 兩種施測方式（紙筆測驗與電腦適性化測驗）的成績結果沒有明顯差異。
2. 在語言及數理能力檢測上兩種施測方式之成績結果沒有明顯差異，但是在英語能力上，針對所有試題類型，電腦化測驗的施測成績結果明顯較高。
3. 電腦化測驗的實驗施測成績與真正施測成績的相關度爲.93，紙筆測驗的實驗成績與真正考試成績的相關度爲.94。
4. 兩種施測方式在實驗施測及真正施測的成績並沒有顯著改變。
5. 兩種施測方式不因性別而有所差異。
6. 兩種施測方式在使用電腦的頻率與施測成績結果之間的關連性皆相似。

此研究以科學公正客觀之方式，驗證適性化測驗施測方式之結果與紙本測驗相同，應考人受測表現不受施測工具不同而有所影響，其結果對於擬發展適性化測驗之相關機構而言，相對地降低其推動適性化測驗之阻力。

五、 美軍軍職性向適性化測驗之改革與創新

會議中上有一篇文章說明美軍軍職性向測驗（Armed Services Vocational Aptitude Battery，ASVAB）發展與推廣適性化測驗之歷程（The Nine Lives of CAT-ASVAB: Innovations and Revelations），美軍軍職性向測驗（ASVAB）每年對百萬名軍職募兵應徵者及高中學生施測，由 ASVAB 評估出來的能力將被運用來決定該名應考人是否可以接受軍職訓練及用來判斷該名應考人可以擔任哪一類軍職工作，此外，ASVAB 評估出來之結果亦幫助學生評估其職能之性向。ASVAB 施測方式是同時間以適性化測驗及紙筆測驗兩種方式施測，大約有 2/3 的募兵應徵者選擇參加電腦適性化測驗（CAT-ASVAB）之施測方式，而美國之長程規劃則擬逐漸以適性化測驗來取代紙筆測驗。CAT-ASVAB 擁有將近 20 年的發展及測驗管理，其優點爲應考人僅需要較少的施測時間、接受較少之試題測試（詳如表 2、表 3），即可以評估應考人之能力與性向，在美國軍職募兵作業上提供相當大的貢獻。

鑑於 CAT-ASVAB 擁有長時間及多版本適性化測驗之發展經驗，此文章對於未來擬發展適性化測驗之機構而言，可以針對 CAT-ASVAB 同時採用紙本與適性化測驗之考量因素、測量程序及題庫初步發展、量表建置的演進、新題庫的發展、操作介面與網路管理等議題，再更深入之考察、觀摩與學習。

六、 新加坡適性化測驗系統

會議中上有一篇文章在簡介新加坡之適性化測驗系統：員工技能系統 Employability Skills System (ESS) （Computerized Adaptive Testing for the Singapore Employability Skills System），ESS 系統主要係檢測新加坡成人工作職場上語言及數理能力之適性化測驗，新加坡透過此系統以提高新加坡求職者及求才者的工作能力與競爭力，並因應新加坡經濟變動的人力需求。

ESS 依據下列程序規則來辦理適性化測驗：選擇一個初始的試題，這個試題大約是題庫的平均中度難易，使用 Rasch 模式來調校及能力估計，並且使用最小標準差來終止一個適性化測驗。

表 2、ASVAB 紙筆測驗及適性測驗施測長度之比較

Test Type	Examinee Ability	Item Difficulty: Easy / Medium / Hard	Test Length
P&P	All	Medium–Hard	30
CAT	Low	Easy–Medium	15
	Medium	Medium	15
	High	Medium–Hard	15

表 3、CAT-ASVAB 測試長度與時間

Subtest*	# of Questions	Time Limit (in Minutes)
General Science (GS)	16	8
Arithmetic Reasoning (AR)	16	39
Word Knowledge (WK)	16	8
Paragraph Comprehension (PC)	11	22
Mathematics Knowledge (MK)	16	20
Electronics Information (EI)	16	8
Auto Information (AI)	11	7
Shop Information (SI)	11	6
Mechanical Comprehension (MC)	16	20
Assembling Objects (AO)	16	16
Total	145	154

在 ESS 的發展過程中，新加坡人力發展部門(Singapore Workforce Development Agency，WDA)對於 ESS 的發展扮演著非常重要的角色，ESS 由 CASAS 設計、發展並客制化，CASAS 公司主要致力於發展成人數學、閱讀及聽力的適性化測驗(CATs)，並且發展寫作及口說的電腦化測驗；而這些適性化測驗是透過安全監控的區域網路及電子存取鑰匙（electronic access key，dongle）來管理安全性。

此文章對於未來擬發展適性化測驗之機構而言，可以針對文章中提及之安全性管控做更深入之探討與學習。

七、 西班牙適性化測驗系統

會議中上有一篇文章(Computerized Adaptive Testing in Spain:Description, Item Parameter Updating, and Future Trends of eCAT)在簡介西班牙之適性化測驗系統：eCAT，eCAT 是一個西班牙國家用來檢測國人英語精熟度的電腦適性化測驗，在同一時間點會有數千位西班牙大學生參加此電腦適性化英文能力測驗。這個測驗是由教育心理學校(Universidad Autónoma de Madrid)的教育心理學者以及知識工程機構（IIC）所合作發展出來的，其合作分工之權責劃分爲教育心理學者負責題庫建置及適性化測驗演算法的設計，而知識工程機構（IIC）負責行銷及控制網頁版適性化測驗的施測管理。

此文章對於未來擬發展適性化測驗之機構而言，可以學習其適性化測驗之發展方式，並將有關適性化專業領域之部分邀請國內外教育心理學者協助完成（如：題庫建置與調校、選題與能力計算之演算法等），至於施測工具（test delivery）之設計與規劃，則可透過資訊委外方式來完成建置與推廣。

八、 印度使用紙筆測驗式 IRT 來發展適性測驗

在會議中尚有一篇有趣的文章，其主旨提到在印度，大規模的測驗是以紙筆測驗(離線)的方式來施測（An Approach to Implementing Adaptive Testing Using Item Response Theory in a Paper-Pencil Mode），印度透過 MeritTrac 幫忙設計試題反應理論(Item Response Theory,IRT)紙筆測驗式考試來分析工科畢業生的能力。

此研究提及在印度要舉辦一個即時線上的測驗系統是很不容易的，因此離線版的適性化測驗對印度來說非常重要，離線版的適性化測驗僅需要一台單一電腦、客製化的學生追蹤軟體、及事前列印好的試卷[1](試卷之試題特徵會依據以往應考人的作答情形與結果，事前計算出試題特徵)即可，印度之作法係透過 1000 位以上應考人來檢測 100 個試題，然後計算出試題的困難度，最後留下來 93 個適合的試題；試題被分爲 6 組(極易、易、中下、中上、難、極難)，每個試卷從這六組中選出 10 個題目(極易 1 題、易 1 題、中下 2 題、中上 2 題、難 2 題、極難 2 題)，同時產生數組 10 題的試卷；而施測方式是以紙筆方式進行，應考人的答案最後會被分類輸入且儲存在一個特別開發的學生追蹤軟體中。

本研究爲一篇突破電腦適性化測驗觀念之研究報告，研究中發現僅包含 6-10 題的試卷與包含 25 題(或以上)的試卷，其檢測效果是不相上下。

[1]試卷：指一次考試所有施測試題之集合。

肆、各國適性化測驗經驗分享

因各國發展適性化之過程與演進對考選部而言極具參考價值，故本報告針對會議中以色列國家提出之適性化測驗經驗分享提出詳細之報告，另鑑於以色列研究報告中對於 GMAC 發展 GMAT 適性化測驗之經驗多有推崇，故本報告亦針對 GMAC 在 2007 年所介紹之 GMAT 發展適性化測驗經驗分享提出詳細之說明。

一、 以色列適性化測驗

(一)摘要

以色列高等教育入學許可所使用的適性化測驗應用於 2 種考試，包含(1)AMIRAM 考試:由許多高等教育機構所使用的英語適性化測驗（the English as a foreign language），這個測驗已經舉辦 22 年，且被使用來做爲學生入學能力分級之工具。(2)入學許可教育心理測驗(Psychometric Entrance Test ,PET)的適性化測驗考試（MIFAM），這個適性化測驗已經舉辦 9 年了，主要是提供給參加高等教育入學許可測驗的身障應考人使用。上述兩種考試（AMIRAM 及 MIFAM）之舉辦皆採用電腦適性化測驗及紙筆測驗兩種施測方式。

在會議中以色列分享之經驗與內容著重於電腦適性化測驗及紙筆測驗兩種施測方式評估能力的等化程序，並且檢測 MIFAM CAT 對於身障應考人施測的適用性，此外，針對入學許可教育心理測驗(PET)轉換到適性化測驗(MIFAM)所遇到的實務問題加以討論，討論之議題包含：(1)內容規範(content specifications) 、(2)試題曝光(item exposure)、(3)題庫(item banks)、(4) 題庫維度(item bank dimensionality)、(5)等化(equating)等議題。

(二)測驗發展機構簡介

以色列國家測驗及評估機構（National Institute for Testing and Evaluation ，NITE）是以色列一個非營利組織，專職眾多高等教育組織之測驗發展、測驗施測、成績評估與報告。NITE 由以色列大學委員會於 1982 年建立，其主辦的第一個教育心理測驗(PET)在 1983 年以紙筆測驗方式舉辦，而接下來以色列相關機構決定採用以試題反應理論(IRT)爲基礎的電腦適性化測驗（CAT），並建立了一個電腦化測驗的單位；

在當時電腦適性化測驗（CAT）是無法廣泛被運用來施測，然而當下之所以有此決定乃鑑於電腦化測驗將會是未來施測的工具，因此以色列進行此項投資。第一個電腦適性化測驗（CAT）是在 1987 年舉辦的，由 NITE 製作供應。

NITE 除了發展電腦化適性測驗外（CAT），也發展了一些電腦化測驗（非適性化測驗），如(1) MEMAD：網頁版的大學預備學校入學及能力分班測驗，一年約有 7,000 位應考人；(2) 入學許可教育心理測驗的語音版本（audio version of PET）：爲視障應考人開發之語音版本，一年約有 230 位應考人；(3) MATAL：主要作爲學習與注意力障礙的測驗評估，到目前爲止約有 2,000 位應考人；(4) 入學許可教育心理測驗考試之網頁版練習程式（Internet based practice tests for PET），於 2008 年有 28,000 位訪客，其中有 4,440 位訪客當年實際參加了教育心理測驗考試；(5)MEMAD 考試之網頁版練習程式，共有 5,040 位訪客，其中有 1,097 位訪客在這九個月中完成了 MEMAD 考試；總體而論，在以色列每年約有 22,000 位應考人採用電腦測驗。

(三)以色列入學許可教育心理測驗（PET）簡介

以色列入學許可教育心理測驗(Psychometric Entrance Test ，PET)是一種學術評量的紙筆測驗，由以色列國家測驗及評估機構（National Institute for Testing and Evaluation ，NITE）發展、建置及管理。所有以色列大專院校採用其測驗評量之結果作爲入學許可之依據。PET 包含 3 大領域：(1)語言推理能力(Verbal Reasoning – V)：共有 60 個試題；(2)數理推理能力(Quantitative Reasoning – Q)：共有 50 個試題；(3) 外語(英文， English as a Foreign Language – E)能力：共有 54 個試題。PET 考試擁有多種語言版本，包含阿拉伯語、俄語、英文、法語及西班牙語，若應考人對於上述語言不熟稔，此考試尚提供希伯來語與英語對照的版本。PET 紙筆測驗之試卷包含 6 個段落、3 個領域（每個領域有包含 2 個段落），每個段落包含 25-30 個試題，作答時間爲 25 分鐘；總成績是以加權方式計分，3 個領域加權比重如下：語言推理能力加權比重爲 2、數理推理能力加權比重爲 2、英語能力加權比重爲 1。

每年約有 80,000 個應考人參加測驗，共計有 60 個機構使用測驗後的結果來作爲入學分班的依據。以色列每年舉辦 5 次 PET 測驗，題庫試題數約 15,000 個試題。

（四）NITE 的適性化測驗

以色列適性化測驗之發展乃基於紙筆測驗之架構，進而發展出適性化測驗施測方式，至於適性化測驗應用於何種測驗乃基於下列原因而選擇 AMIRAM 及 MIFAM

兩種考試來辦理適性化測驗。

1. 以色列不論是在地理範圍或是人口數上皆屬於少數，而應考人數量的多寡將限制新試卷預試及組成試題母版試卷的數量，試題母版試卷數量將影響每年可以舉辦考試的次數、每次考試的應考人數及每次考試需準備的考場、試區、電腦座位數等，而這些將影響舉辦考試的總成本。
2. 入學許可教育心理測驗(PET)是一種高利害關係的入學許可測驗，測驗結果提供給以色列所有大學作爲入學許可之依據，因此，在以色列有關此種考試的補習風氣旺盛，約 80%的應考人會參加補習班考前準備課程。對於 NITE 來說如何降低試卷重複使用率是非常重要的；尤爲甚者；近來以色列國會立法要求考後須公布考畢試題。基於上述原因，爲確保適性化測驗（CAT）安全性，NITE 必須每年發展至少 12 個紙筆測驗的試卷。
3. 以色列高等教育入學許可考試有多種語言之版本，希伯來語的試卷必須翻譯成 5 種語言(就阿拉伯語的測驗而言，每年舉辦了 4 次)。大體而言，語言推理能力試卷產製過程中，必須提供 2 份希伯來語的試卷方可以產生 1 份阿拉伯語的試卷；也就是說，每年必須提供 8 份希伯來語的試卷來辦理 4 次阿拉伯語的測驗。在發展不同語言版本試卷時，對於希伯來語紙本試卷數量要求不低。
4. NITE 的預算是由測驗的報名費用而來，從相關文獻與 NITE 的財務預估中發現，若要建置一個有效率且安全性的電腦測驗系統，NITE 必須有更多的財務支援。

基於上述安全性及成本考量之因素，以色列在持續發展 CAT 的架構時，將對 CAT 適用的應考人及種類有所規範，僅將 CAT 應用於兩種考試：(1)AMIRAM，PET 的外語能力領域；(2) MIFAM，PET 的適性化測驗，提供給參與 PET 測驗的身障應考人使用。這兩種測驗非僅單一使用適性化測驗，而是同時提供紙筆測驗與適性測驗兩種方式供應考人選擇。

(五)MIFAM 概觀

MIFAM 是 PET 的適性化測驗版本，提供給參與 PET 測驗的身障應考人使用。其發展的演進係因爲在過去的幾年中，申請就讀大學的學生對於大學入學考試逐漸有特殊的應考需求，且這個需求有逐漸成長的趨勢；在 2008 年，約有 5%的大學入學申請者要求特殊的應考設施與規則，要求者中有 65%符合特殊的應考規則，因而引起一些相關議題的討論，諸如：對於身障學生測驗的公平性、特殊與標準化考試的比較等。

基於大學入學申請者要求提供學習障礙應考人及生理障礙應考人特殊應試測驗服務，以色列衡酌測驗公平性與標準化等議題，因此 NITE 發展了 PET 的適性化測驗版本（MIFAM）。這個適性化測驗對於「每個試題」都配置有應考的作答時間，且符合身障應考人特殊的應考需求，並且可以與標準測驗相互比較。不過，雖然 MIFAM 大部分係提供給身障應考人使用，但是其設計原則卻是針對所有應考人而設計，非僅單純考量身障應考人。此外，MIFAM 也解決之前紙筆測驗中相同考試採用非標準化條件之問題：即 MIFAM 爲適性化測驗，所以應考人僅需要回答較少的題目即可評估應考人之能力，雖然賦予身障應考人每題較多的作答時間，但因爲作答題目減少，故解決身障應考人採用紙筆測驗時所延長過多應考時間之問題。

MIFAM 是一種高度利害關係的測驗，因此較有可能會有應考人鋌而走險，此測驗在安全性上的議題上要求較高。

在發展 MIFAM 之前，NITE 針對有學習障礙及正常學生辦理學生意見調查，調查結果發現兩種學生對於電腦的態度沒有明顯的差異性。第一次 MIFAM 在 2000 年 7 月舉辦，在 2008 年有 1,000 位應考人採用 MIFAM 應考，截至目前爲止，已經有 5,100 位應考人採用 MIFAM 應考，且透由 MIFAM 應考人的問卷調查回覆中發現，應考人對於 MIFAM 的適性化測驗之特色很滿意，並且認爲 MIFAM 是清楚且友善的，此外，對於閱讀速度慢與注意力分散的應考人而言，每題分配的作答時間是非常友善的，這些應考人認同考試公平性。

透過 MIFAM 實際測驗後的資料顯示，採用 CAT 評估身障應考人能力之施測方式，其測驗評估能力之精確度等同於身障應考人採用紙筆測驗延長考試時間所評估之能力，甚至於較紙筆測驗更爲精確地評估出應考人之能力值。

(六)AMIRAM 概觀

以色列入學許可教育心理測驗（PET）包含三大領域：語言推理能力測驗（V）、數理推理能力測驗（Q）、外語（英語）能力測驗（E），其中英語能力測驗（E）有兩個作用：其一，這是 PET 的一個子測驗，其二，這個測驗結果可以用來作爲學生外語（英語）能力分班的一種依據。AMIRAM 是外語（英語）能力測驗（E）的適性化測驗版本，主要是作爲能力分班的依據，這個測驗包含三類問題：句子完成、句子重組、閱讀測驗。每年約有 12,000 個應考人參加這個測驗，截至目前爲止約有 117,000 個應考人參加這個測驗，這是一種中度利害關係的測驗，因此應考人較不可能鋌而走險，此測驗對於安全性的議題上要求較不高。

(七)適性化測驗建置之議題

NITE 的適性化測驗採用三參數試題反應理論模式（3-parameter (3-PL) item response theory model），NITE 開發一個 NITECATSYS 軟體來建置適性化測驗，NITECATSYS 之程式提供測驗產生、品質保證及測驗施測之功能，此外，尚包含人機介面（HMI）及管理功能之額外模組。至於適性化測驗之試題則挑選自紙筆測驗考試。以色列電腦適性化測驗建置過程包含下列之特色：

1. 單向度（Unidimensionality）

   入學許可教育心理測驗（PET）包含三大領域：語言推理能力測驗（V）、數理推理能力測驗（Q）、外語（英語）能力測驗，建置適性化測驗的第一步驟即是調查每一個領域的單向度，確定這三大領域符合試題反應理論(IRT)的基礎假設需求。

2. 參數估計（Parameter Estimation）

   參數估計是依據紙筆測驗實際資料來估算，原本參數估計是採用 Assessment System Corporation (1987)ASCAL 的軟體來估算；後來 NITE 爲了同時處理大量試題自行開發一個參數估算軟體 NITEST；2002 之後，參數估計採用 means of Bilog-MG。

3. 測驗架構（Test Structure）

   以色列適性化測驗之起始選題策略爲：在每個領域的前兩個試題採用中難易度及低鑑別度的試題，這些試題是隨機抽樣的，這種初步選題的方式乃基於前兩題不應該太困難，也不應該太簡單，且後續試題之

變異數亦不能有太大的差異。至於以色列適性化測驗之終止規則爲：當適性化測驗達到下列兩者其一之標準時，適性化測驗即停止：(1)接續的變異數小於預設值，且應考人已完成最少應作答題數；(2)應考人完成最大需作答試題數。

4. 有關以色列適性化測驗 CAT 之架構及每個試題類型的時間分配詳如表 4，應考人作答原則爲：每個應考人不能回到前一題作答、回答過的試題不能修改作答選項、不能跳答試題。不過應考人是被允許在分配作答的時間範圍內不作答。以色列適性化測驗軟體允許不同的選題規則，例如依據最大資訊量（maximum information）、試題的困難度（difficulty of the item）、最大資訊量與試題困難度的結合模式、隨機選題或是循序選題。

表 4、以色列適性化測驗 CAT 之架構及每個試題類型的時間分配

Domain	Item Type	P&P	CAT	Time Allotment per Item (in Min.)
Verbal Reasoning	Words and phrases	~13	10-15	1.0
	Verbal analogies	~20	13-31	1.5
	Letter Switching	~13	12-16	3.0
	Sentence Completions	~17	17-23	3.0
	Logic	~17	17-23	4.0
	Reading Comprehension	~20	12-16	7.0 (per text), 4 (per item)
Total Computerized				100.0-118.0
Total P&P				50.0
Quantitative Reasoning	Questions and Problems	~60	57-66	4.0
	Diagrams & tables	~16	11-14	5.0 (per graph), 4.0 (per item)
	Quantitative Comparison	~24	23-29	4.0
Total Computerized				117.0-157.0
Total P&P				50.0
English	Sentence Completions	~41	38-54	2.0
	Restatements	~22	29-38	4.0
	Reading Comprehension	~37	18-24	7.0 (per text), 4.0 (per item)
Total Computerized				75.0-89.0
Total P&P				50.0

5. 在正式使用 CAT 做爲施測工具之前，必須模擬適性化施測結果以確保某些試題之曝光率不會太高、某些試題曝光率不會太低，且須確保多數應考人可以達到一個事前決定的最小變異數。以色列試題曝光率的驗證程序是以試題類型爲單位，對某一個試題類型而言，若是該試題類型的試題曝光率過低，其解決方式則是增加該類型試題被挑選的機會；反之，若是某一類試題之曝光率過高，則其解決方式就是降低該類試題被挑選的機率，或是增加該類試題之試題數量；此外，改變選題規則也是方法之一。舉例來說，假設某一類試題的試題組共有 10 個試題，試題 1、試題 2、試題 3、試題 4、…試題 10，而這 10 個試題的其中兩個試題會被挑選出來作爲考試試題，若試題 10 經由模擬程序顯示其試題曝光率過高，則可以將這 10 個試題分爲兩組，分別爲試題 1 至試題 9 一組、試題 1 至試題 10 一組，這樣，就可以降低試題 10 的試題曝光率。
6. 後變異數（Posterior variance）被視爲檢驗能力水準的一種函數，如果模擬資料顯示具有一定能力水準之應考人中，有相當高的比例具有高度後變異數，則題庫需要再新增一些高難度的試題，當然以色列也可能會給應考人增加應考試題或是在題庫中增加某一困難度的試題，有時，改變應考人應考時所看到的試題類型順序也是有必要的。不過值得注意的是，有一些試題是比較容易曝光且容易被記憶的，針對這樣的試題需要更小心處理。
7. 內容規範（Content Specifications）

   因爲 PET 的兩種施測方式（紙筆測驗與適性化測驗）是同時施測，因此 NITE 決定電腦化測驗的內容規範將遵循紙筆測驗的要求，且題庫的管理是依據試題類型做爲管理單位。

   以色列每年發展的適性化測驗試卷數量受限於實際紙筆測驗考試所發展的試卷數量，一個適性化測驗試卷是由 3 或 4 份紙筆試卷所組成，此外再依據需求額外新增一些試題，每一個試題題組（item pool）包含之試題數如下：語言推理能力試題（V）210 題、數理推理能力試題（Q）175 題、外語（英語）能力試題（E）230 題，這種試題題組（item

pool）試題數之規範乃基於 Runder(2007)GMAC 的研究結果而設計的，此外，以色列也透過模擬測驗來確保新的試卷擁有良好的收斂度及曝光率。

(八)適性化測驗的特色

以色列 PET 電腦化測驗施測方式與紙筆測驗施測方式最大的不同點就是針對每一個試題皆有作答時間限制，這個特色也與其他電腦化測驗不相同，而這樣的特色可以讓其標準化作業以單一試題爲單位。在以色列的適性化測驗中，相較於紙筆測驗施測方式，應考人僅需作答一半的試題數，而時間的分配則爲紙筆測驗的 100%到 400%。採用適性化測驗之優點就是僅需作答較少的試題、針對每一試題分配作答時間、使用友善的作答介面（大字型、獨立顯示每一個試題、休息時應試時間暫停等）。有關以色列 CAT 測驗的架構（測試的三大領域、試題類型、每一試題類型試題數、每個試題類型的時間分配）詳如表 4，特別需提醒的是實際上適性化測驗應考人的考試時間會較表列時間爲短。

(九)適性化測驗量尺的轉換

基於應考人可以自由選擇以紙筆施測方式或適性化施測方式來參加考試，因此主辦考試單位必須保證兩種施測方式在衡量應考人能力表現的量尺上具備同等性（兩個量尺是相同的）。紙筆測驗與適性化測驗兩種施測方式的成績校準程序有三個步驟：(1)適性化測驗測得的能力值轉換成相對的答對題數得分（number-correct score）；(2) 答對題數得分（number-correct score）轉換成標準化成績；(3)修正標準化之成績。

NITE 以實驗來比較適性化測驗的評估能力，首先，針對報名應考的應考人隨機抽樣邀請其參加實驗測驗，並依據其實驗時的估計分數，與後來真正考試的分數進行比較，實驗發現(1)應考人在適性化測驗所得到的成績與實驗測試時的紙筆測驗成績、真正紙筆測驗時所得到的成績相似；(2)適性化測驗評估應考人能力與紙筆測驗評估結果相似；(3)抽樣實驗的相關係數（實驗與真實考試結果的相關係數）與真實應考人前後兩次考試的相關係數相似。

此外，以色列還比較「性別」、「電腦的使用經驗」與對不同施測方式（紙筆測驗與適性化測驗）的影響，實驗結果發現性別對於不同施測方式的影響不顯著，同樣地電腦的使用經驗對於不同施測方式的影響也不顯著。

(十)以色列適性化測驗發展演進

以色列每一個 MIFAM/AMIRAM 的適性化測驗版本在其開發、建置、測試與事後管理階段皆有詳細的驗證過程，而這些驗證過程可以作爲台灣未來評估、發展適性化測驗之參考。以下針對 AMIRAM 最新一版適性化測驗的發展過程說明如下：

發展階段： 在發展新一代版本時（尚未正式施測之前），需做好品質控管，並儘速的完成下列各項檢核程序：

1. 選擇 3 份英語能力測驗（E）的紙筆測驗試卷及新一版適性化測驗所需要的其他額外試題，定義好相關的限制與要求。
2. 檢核所有試題類型的資訊函數，並與前一個版本相比較。
3. 確認試題參數的正確性，並應與系統相一致。
4. 檢核適性化測驗能力估算值轉換至標準分數的轉換表格，並與前一版本相比較。
5. 準備安裝之軟體。
6. 依據試題類型決定試題呈現之順序，例如，句子重組試題類型通常在閱讀測驗類型之前，因爲閱讀測驗之試題通常會是好幾題組成的題組，因此適性化的程序（隨機選出下一試題）將會受限，這個議題尤其當能力估算已經接近尾聲時特別引起關注；而句子重組試題通常可以隨機地選擇單一試題，因此較適用於適性化隨機選取。
7. 決定適性化測驗結束的條件，近來常使用之適性化測驗終止規則爲：當應考人的能力估算值超過某些切割點（高於或低於切割點，即使後變異數並未達到事前定義的預設值）即終止考試。此種適性化測驗終止的規則通常對於能力分班非常有幫助，且應考人通常在回答 23 個試題後即會結束考試，對於應考人的能力值較爲極端時，是不需要多回答一些額外不相關的試題，且也可以降低試題的曝光率。
8. 針對每一個試題做最後的檢查，包含關鍵文字及指令（key and instructions）。
9. 至少要求兩組不同工作人員預估施測及評分階段可能遇到的各種問題，包含時間分配問題、評分問題、考試各階段所需要之復原程序等。
10. 辦理模擬實驗，針對下列問題須透過模擬實驗加以驗證：

11. 量尺轉換後標準化成績的分佈情形，須與前一版本相比較。
12. 後變異數小於 0.08 的應考人比率（期望值爲 0.92）
13. 能力估算值的平均數與標準差，真實能力值與後變異數值。
14. 測試長度（指應考人應考試題數）的範圍及平均值。
15. 能力估算值與真實能力值之相關係數。
16. 全對、兩個錯誤或是全錯的標準分數值。
17. 確認適性化測驗考試係依據終止條件而終止。
18. 確認每一試題在不同標準下之試題曝光率。
19. 在可攜式電腦硬碟上刪除前一版本。
20. 準備備份硬碟。
21. 針對新版本記載所有的改變及決定。

施測階段：施測時須做好適性化測驗之品質控管，針對下列議題逐項執行與檢視

1. 自資料庫中擷取資料。
2. 檢查適性化測驗辦理的次數。
3. 檢查不同變數的能力值分佈。
4. 檢查能力值與答對的試題數（比率）。
5. 檢查答錯的試題數及剔除的成績。
6. 檢查每一個應考人的應試試題數目。
7. 檢查達到 Pvar 或未達到 Pvar 的應考人比率。
8. 依據應考人 Pvar 檢查應考人的應試試題數。
9. 檢查從未被使用的試題。
10. 比較模擬考試及真正考試之試題曝光率。
11. 找出不正常應試時間之應考人。

(十一)結論

本文章係探討以色列在設計、發展適性化測驗考試（MIFAM / AMIRAM）不同版本時所考量的實際議題。其中一些重要的考量項目說明如下：

1. 爲了要降低安全風險，因此以色列僅應用適性化測驗在兩種考試上，一種是給身障應考人使用的 PET 適性化版本（應考人數不多），另一個是外語（英語）能力測驗（爲非高利害關係之考試）。
2. 以色列同時間提供應考人紙筆測驗及適性化測驗兩種施測方式，此兩種施測方式的內容規範是相同的，且能力評估結果可以等化。截至目前爲止以色列發現兩種施測方式的評估能力是相同的。
3. 題庫特徵：基於安全考量，以色列 NITE 決定建置多個小型的題庫，而不是一個大的組合題庫，每一個小題庫是由 2 到 3 個紙筆測驗試卷所形成，因此與紙筆測驗有相同的內容規範。
4. 適性化測驗演算法：爲了有最大正確率及控制試題曝光率，以色列運用了多種試題選擇的參數，而最後試題選擇演算法則是由模擬實驗之結果所決定。
5. 時間分配與應試長度：不同於其他的適性化測驗系統，以色列適性化測驗係針對每一個試題分配應試時間限制，並且有不同的考試長度。
6. 品質控管：因爲適性化測驗有時給人黑箱作業的印象，因此品質控管非常重要，以色列嚴格做好品質控管，經由量化分析確認適性化測驗可以真正評估應考人之真實能力。
7. 適宜性：適性化測驗須確認施測結果是否適宜，以色列兩個適性化測驗被驗證是有效率並有價值的，並可滿足部分應考人的特殊應考需求。
8. 滿意度：適性化測驗須檢測是否符合建置目標，以色列應考人對於這兩個適性化測驗反映出高度的滿意度。

二、 GMAT 適性化測驗

(一)前言

美國商學研究生入學測驗（GMAT）在 2007 分享之 GMAT 適性化測驗經驗分享時提及：一個成功的適性化測驗主要是對於實務問題是否有深入的研究與探討，在適性化測驗設計與規劃時，測驗主管單位應著重於內容規範（test specifications）、試題挑選演算法（item selection algorithms）、題庫設計（item bank design and rotation）、能力估算（ability estimation）、前測（pre-testing）、試題分析（item analysis）、資料庫設計（database design）及資料安全（data security）等議題，且適性化測驗開發廠商須落實適性化測驗之建置目標。

美國研究生入學管理委員會（GMAC）建置適性化測驗已有 10 年經驗，其主導之 GMAT 考試是一個美國商業及管理學院針對申請入學學生之能力評估測驗，GMAT 包含三大部分：(1)分析寫作評估（Analytical Writing Assessment，AWA）；(2)數理能力評估（Quantitative section）；(3)語言能力評估（Verbal section），每年約有 20 萬應考人參加這個考試，其評估結果提供個 3000 個不同程式作分析，應考人須先報名預約考試，而全球約有 400 個測試中心。GMAT 在適性化測驗之建置上採用 3 參數試題反應模式（3-parameter (3-PL) item response theory model），試題依據試題參數校準與評估，題庫依據 3 參數試題反應模式來建立，且適性化測驗演算法採用 3 參數來挑選試題。

有關 GMAT 測驗之內容規範，表 5 提供 GMAT 測驗內容單元、分配作答時間及成績配分，總測驗時間是 2.5 小時，不包含一個簡短的問卷時間及休息時間，雖然 GMAT 包含的內容向度可能跟一般入學許可測驗考試相似，但是 GMAT 強調商業的概念，著重在語言及數理的邏輯推理，並且使用與商業相關的內容來施測。

舉例來說，在資訊量判別（data sufficiency）這個測驗單元中，應考人需要決定試題提供的資訊是否足夠解決試題所描述的問題（應考人並不會被要求解決這個問題），這類的問題主要是評估應考人分析計量問題的能力及應考人判別資訊相關度與資訊量是否足夠之能力，資訊量判別範例題型詳如圖 1。

GMAT 數理能力的內容規範基本上要求須涵蓋算數、代數與幾何三個基本的能力測驗，此外，也有一些試題是與數學及數學公式有關的，GMAT 在數理試題內容

規範上要求試題內容基本上應屬於算數、代數與幾何這三大基本技能之一，而對於一些特殊技能則定義試題數之最大上限（舉例來說，對於評估三角函數及百分比技能之試題，GMAT 會限制其出現在試卷中之上限比例），此外，對於與性別相關之內容及整體試卷正確答案出現的機率（如整體試卷正確答案爲 A 的機率）亦有出題題數之上、下限之限制。總體而論，GMAT 的數理能力評估有 27 個內容規範限制，語言能力評估有更多的內容規範限制，以圖 1 爲例，此試題即包含下列的內容規範：爲資訊量判別、代數及百分比能力評估之試題，正確答案座落於 D，試題與性別無關。

表 5、GMAT 測驗內容單元、分配作答時間及成績配分

GMAT® Section	Number of Questions	Allotted Time	Scoring
Analytical Writing Assessment Analysis of an Issue Analysis of an Argument	 1 1	60 minutes 30 minutes 30 minutes	0 – 6 (half-point increments)
Quantitative Problem Solving Data Sufficiency	37	75 minutes	0 – 60 (1-point increments)
Verbal Sentence Correction Critical Reasoning Reading Comprehension	41	75 minutes	0 – 60 (1-point increments)

If a real estate agent received a commission of 6 percent of the selling price of a certain house, what was the selling price of the house?

(1) The selling price minus the real estate agent's commission was $84,600.
(2) The selling price was 250 percent of the original purchase price of $36,000.

(A) Statement (1) ALONE is sufficient, but statement (2) alone is not sufficient.

(B) Statement (2) ALONE is sufficient, but statement (1) alone is not sufficient.

(C) BOTH statements TOGETHER are sufficient, but NEITHER statement ALONE is sufficient.

(D) EACH statement ALONE is sufficient.

(E) Statements (1) and (2) TOGETHER are NOT sufficient.

The correct answer is D.

圖 1、GMAT 試題與內容規範範例

(二)採用適性化測驗之歷史演進

GMAT 在 1997 年 10 月開始辦理第一次適性化測驗，在當時 GMAC 推動適性化測驗主要原因有兩個：(1)辦理適性化測驗是爲了舉辦考試之方便性：GMAT 每年舉辦四次，近來報考之應考人數量逐年增加，但是卻越來越難取得應試座位，應試座位數難取得之狀況對美國以外區域之應考人尤爲甚之；(2)GMAC 推動適性化測驗另一個議題是因爲有越來越多入學考試應考人在 GMAT 的成績得分很高，然而許多學校對於這些高得分應考人越來越難區分其差異性。

1992 年，美國教育測驗服務機構（Educational Testing Service ，ETS)向 GMAC 董事會首次介紹適性化測驗，在當時 ETS 正積極地將其公司數個客戶（包含 GRE 及 TOEFL）導向電腦適性化測驗，ETS 此舉主要係爲了增加採用適性化測驗之客戶數量以提高其施測與管理之經濟效益。ETS 對於 GMAC 董事會的介紹中提及採用適性化測驗可以增加舉辦考試的便利性、使用新試題類型、並且在未來可以新增新的評估測驗。

1993 年，GMAC 董事會接受第一次適性化測驗正式的展示，此展示介紹 GMAT 轉換到適性化測驗的潛在利益，包含提供更多辦理全球考試的次數，並且可以有效區別高得分應考人之能力，至於轉換成本則包含修改應考人註冊報名系統、題庫系統、成績表報系統等，因爲 GMAC 已經有一個相當大的題庫系統，因此並未將增加試題包含於費用評估中，整個轉換的費用約需花費美金 400 至 700 萬。

GMAC 董事會於 1995 年通過採用適性化測驗，且開始跟 GMAC 成員及其他使用 GMAT 評估結果的使用者介紹這個計畫，因爲 GMAC 並沒有專職的教育心理測驗專家，因此一個獨立的第三者（顧問）將負責建議 GMAC 這個轉換計畫的價值，並且監督由紙筆測驗轉換至電腦化測驗的移轉過程，而這個顧問的主要貢獻之一即是堅持必須比較適性化測驗與紙筆測驗之結果。

1996 中期，當 GMAC 已告知他的客戶這個轉換的利益及所需額外的費用後，ETS 發現費用評估過程中少估算了發展新試題的額外費用，並且將此問題告知董事會，GMAC 在此時已經決定建置 GMAT 適性化測驗，因此，這個額外的插曲對 GMAC 來說是一個很大的風險，而 GMAT 整個最後轉換的費用，包含發展新試題、整體個建設等項目之建置費用將近美元 1,170 萬。

1996 年 10 月，在正式辦理適性化測驗的 12 個月前，一些有關電腦適性化測驗與紙筆測驗的比較性研究開始進行，第一個研究是邀請報名參與考試的應考人來參加這個實驗研究，參加實驗者必須完成適性化測驗與紙筆測驗（隨機指派兩種施測方式的順序）。此研究共邀請 10,196 位報名者參與研究，其中 4,300 位報名者願意參加實驗，參加實驗者中，3,606 位完成了適性化測驗考試，而 2,545 位完成了適性化測驗及紙筆測驗考試。在可用的樣本中，比較先參加紙筆測驗的實驗結果與先參加適性化測驗的實驗結果，發現在許多重要的衡量變數中差異性很大，而所有的實驗結果與真正紙筆測驗考試的結果也有很大差別。這個研究發現紙筆測驗與電腦化測驗是無法相比較的，且必須有一個很大的等化調整程序。這個問題的原因之一是因爲適性化測驗的要求作答的速度太快，約有 18%的應考人來不及回答最後兩個數理推理試題，爲了解決這個問題，ETC 決定增加數理推理單元 5 分鐘中的適性化測驗作答時間，並且減少兩個作答試題數。

第二個研究在 1997 年 4 月提出（在正式辦理電腦化測驗的前六個月），因爲時間限制，因此實驗的方式爲實驗者皆先參加紙筆測驗，此實驗邀請 3,000 位註冊者參加實驗，其中僅有 773 位參加紙筆測驗後也參加電腦化適性測驗。很明顯地，許多應考人對於他們的紙筆測驗成績很滿意，因此沒有繼續參加電腦化適性測驗。此實驗因爲樣本數太少的關係，並無法進行等化研究，因此最後等化研究的資料組合了 1996 年 10 月及 1997 年 4 月兩次的實驗資料來進行等化研究。

在 1996 年 10 月的這個實驗平均 GMAT 成績偏高，在 1997 年 4 月的這個實驗平均 GMAT 成績偏低，不過，在 1997 年 4 月這個先舉辦紙筆測驗的實驗中，因參與實驗之樣本太少而不具代表性。

從 1996 年及 1997 年的研究中發現下列之問題：(1)適性化測驗的量尺與紙筆測驗的測驗成績不完全一致；(2)適性化測驗平均數理理解能力之成績明顯較高；(3)此時的適性化測驗無法有效分別區高得分群之能力。

值得注意的是 10 年之後，GMAC 換了另一個合作廠商（由 ACT 設計考題，而由 Pearson/VUE 提供考試服務），也運用了另一個評估等化的方法，這個方法克服了原本 1996 年的研究問題，GMAC 及入學許可的官員非常滿意這個結果，因爲原本使用適性化測驗的主要目的可以達到了，使用紙筆測驗的成績與適性化測驗的成績相同，而且，侷限於試場座位數無法因應越來越多應考人之問題也改善了。

(三)建置時考量之議題

GMAC 在建置電腦適性化測驗時，考慮下列之議題詳列臚下：

1. 須符合內容規範

   內容規範定義了整個測驗要衡量之項目與內容，因此適性化測驗是否符合內容規範相當的重要，這個議題包含了如何從廣大的題庫中挑選試題，且被挑選的試題必須符合測驗中所要求的各種內容規範。以 GMAT 爲例，GMAT 對於內容規範所採用的方法是區別每一個試題個別的規範及整個題庫的規範，舉例來說，數理理解能力的試題須包含 3 個領域/7 種題型類別，其題型分別爲：技能領域（包含資訊量判別及問題解決 2 種題型類別）、內容領域（包含代數、運算與幾何 3 種題型類別）及應用領域（包含應用及公式導向題型類別），GMAC 要求題庫中的試題必須針對這 7 種題型類別各自包含一定數量的試題，但是不限制題庫中混和題型類別（一個試題含有多個題型類別特徵，如一試題同時屬於資訊量判別、代數及應用題型）之試題數量，針對上述內容規範的要求，GMAC 的作法係將每一個試題事前定義其內容屬於何種題型類別。此外，針對整個題庫尚包含較不重要的規範要求，舉例來說，題庫中對於試題正解出現的位置要求必須平均分配，不可以全部落於答案 A（或答案 B、答案 C、答案 D），試題內容亦不不可偏重某一性別，試題內容必須符合測驗之主題內容或是其他的測試特徵。

2. 試題使用、曝光及適性化演算法

   發展試題是非常昂貴的，通常一個試題大約需花費 1,500 美金至 2,500 美金，鑑於發展試題非常昂貴，因此測驗機構非常在意是否所有的試題皆被使用過，有無過度使用之試題。

   以 GMAT 爲例，圖 2 顯示 GMAT 以往基於最大資訊演算法，其實際題庫中的試題曝光率分佈情形，約有 28%的試題從未被使用過，18%的試題出現在 15%的應考人試卷中。因此，GMAC 在研究中建議，若使用最大資訊演算法來挑選試題時，必須再輔佐試題曝光率控制，否則會有一些試題被過度使用，而有一些試題則很少被使用。此外，還有一個問題會出現於考試開始之時間點：適性化測驗試題之挑選是依

據前一個試題回答結果所估算之能力值來挑選下一個試題，而在考試剛開始時，這個應考人初估的能力值往往不是很準確的，所以被挑選出來的試題困難度通常與應考人真實能力遠不相符，因此，在考試一開始即運用最大資訊演算法來挑選一個提供最大資訊、最具鑑別度試題的作法是很浪費資源的，且一些好的試題可能在考試一開始就曝露給應考人，但此時對於鑑別應考人真實能力卻鮮少有幫助。

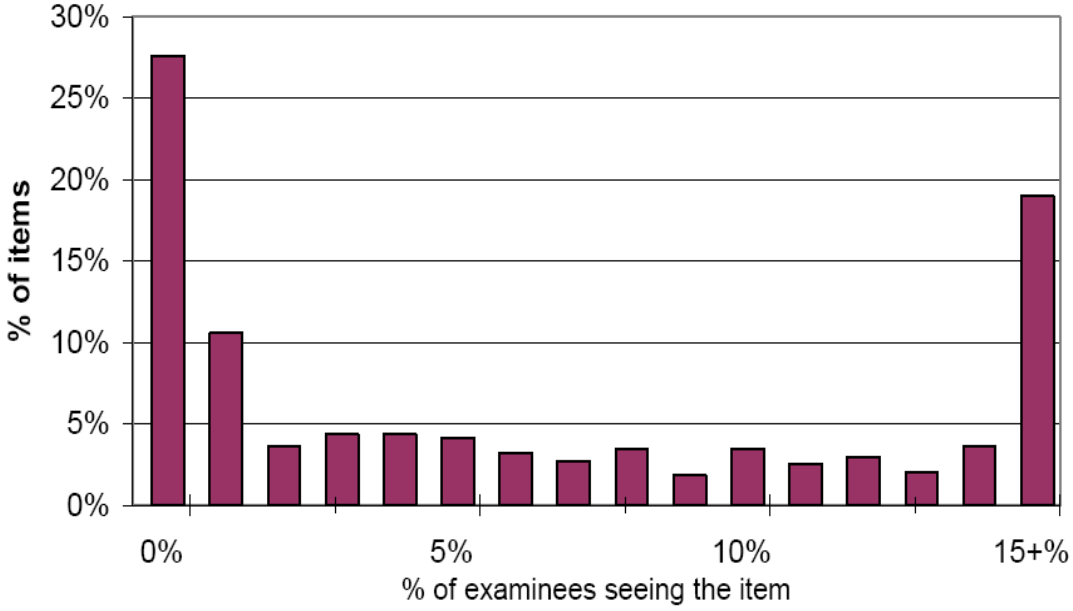


圖 2、GMAT 基於最大資訊演算法，題庫試題曝光率分佈情形

就 GMAC 建置 GMAT 適性化測驗之實際經驗，其試題曝光風險的衡量是計算相同試題呈現給相似能力應考人的機率。若應考人數固定，則題庫試題數越大，試題曝光率越低。此外，GMAT 降低條件式試題曝光率的方法是同時間從一些相同內容範圍的題庫中隨機挑選試題，且在不同區域使用不同的題庫。

3. 題庫特色

GMAC 在 1997 年將紙筆測驗轉換至適性化測驗的準備過程中，GMAC 建置了超過 9,000 個具品質的試題，且從此 GMAC 穩定地增加試題數，而對 GMAC 來說，建置試題的一個挑戰就是如何將一個題庫（item

bank）切割成數個試題題組（item pools），而這些試題題組(item pools)可以符合規範並且讓所有應考人被分配到的試題皆滿足標準誤差值。理想的適性化測驗試題題組（item pool）係指包含一堆高鑑別度的試題，這些試題涵蓋著每一能力階層的試題內容，而且此類試題的資訊函數圖形即是橫跨在各能力值的一連串高峰分配圖形。

試題題組的形成（pool formation）方法之一是將題庫中所有可用試題放入此題組中，當然，這種方法會保證題組中包含最好的可用試題，然而，這樣的方法卻讓試題暴露在試題洩題之安全危機中。身爲一個測驗的管理機構，GMAC 期望的方式是希望能夠盡可能的使用最小的題組，且這個最小的題組卻能包含所需要的試題內容規範。Weiss(1985)指出一個令人滿意的適性化測驗需要一個包含 100 個高品質試題數的題組，而這 100 個試題是分佈在各能力階層的；當然，若有一個包含 150-200 建構良好的題組則更好；若一個測驗對應考人而言有高度利害關係（測驗通過與否對應考人來說很重要），或這各個測驗有許多限制（如試題選擇之原則是隨機挑選最大試題資訊之試題，且必須符合最少試題曝光率，或是必須符合多項試題內容規範）時，則就需要一個較多試題數目的試題題組，以 GMAT 爲例，因爲 GMAT 有許多內容規範與限制，且有目標標準差之要求，因此試題題組的數量約爲 600 題試題到 1,000 題試題不等。GMAT 試題題組(item pool)通常儘量不包含相似的試題，且亦不包含最近使用過之題組試題（試題被使用過後需要休息一陣子才會再度被使用）。

GMAT 對於試題題組(item pool)品質之控制方式係重複地以模擬的方式評估試題曝光率期望值、試題題組覆蓋率及條件標準誤差期望值（expected conditional standard errors），並透過置換試題及新增試題方式調整條件標準誤差值，直到其符合預定水平。目前這個模擬程序由 GMAC 的測驗發展廠商(ACT)針對每個試題題組（item pool）來執行，且由 GMAC 每個月檢視模擬結果，並偶爾將模擬資料與實際資料進行比對。此外值得一提的是，因爲有許多補習業者或其他人會想盡辦法拿到考過的試題，因此 GMAC 定義了一個重複使用以前試題題組的最

大比率，以降低重複使用試題題組的風險，此外，GMAC 現在有超過兩個以上全職員工負責來監控各種與 GMAC 相關的侵權行爲。

4. 試題偏差

檢測試題函數差異性的一個常用方法是透過不同群組來校準試題參數，GMAC 在試題前測評估階段即以此方法定期來檢測試題函數的差異性。

不過前測資料通常受限於各群組的受測人數（the number of examinees in subgroups），因此，GMAC 在檢測試題偏差時也使用實際的試題，舉例來說，GMAC 對於 GMAT 試題是否對於歐洲應考人有試題偏差之問題極有興趣，因此，Guo, Rudner, Owens, and Talento-Miller (2006) 提出了有效的方法，可以利用適性化測驗實際考試的資料來研究試題偏差，從次群組反應中所估算出來的試題反應函數（IRF）會被拿來與實際考試時試題參數所定義的試題反應函數(IRF)相比較，若發現在次群組中之應考人並無法如同正式考試應考人般擁有相同正確作答的機率時（兩組相比較之應考人具有相同能力），即代表有試題偏差。這個方法在試題前測，試題校準及修改試題參數上，皆有很好的效果。

5. 試題參數修正

當試題被校準過且擁有相當品質後，卻可能因爲一些理由而淘汰這個試題，因此，我們必須考量到個別試題參數在校準的標準誤差之下是否會因時間而改變，此外並考量這個試題參數的改變是否會有其他影響。GMAC 已成功地在同時間重新校準整個試題題組之試題反應資料，包含非正式考試之試題，而這個方法確保了試題參數的穩定度。

(四)結論

一個成功適性化測驗的關鍵因素是該測驗系統是否可以精確地評估應考人的能力值，且將試題曝光率及試題風險降至最低。GMAC 提供了一些設計及評估美國商學研究生入學測驗（GMAT）適性化測驗時所考量的一些實際議題，這些議題非常值得在建置適性化測驗時引用參考：

1. 內容規範：當考量眾多試題內容規範的平衡點時，應該保證每一位應考人的內容規範要相似，GMAT 的內容規範定義每一個被挑選出來給應考人作答的試題需要具備的內容與條件，且規定了題庫的基本需求、條件誤差及其信度。
2. 試題使用、曝光及適性化演算法：大部分研究試題曝光的議題多著墨在過渡曝光，有品質的試題是非常昂貴地，通常測驗的主辦單位非常在意試題的投資是否有最大的使用效益（是否每一個試題皆被使用過），GMAC 要求相同的試題不可以出現在太多應考人的試卷中，同時，依據最大資訊（maximum information）的演算法來控制試題使用及試題曝光。
3. 題庫特徵：雖然將所有可用的試題置於一個測驗試題題組(test pool)中有其教育心理測驗的優點，不過實務上之考量重點卻應該著重於安全性議題。以測驗主辦單位的角度來看，小型的測驗試題題組(small pool)非常吸引人，不過卻會有條件式試題曝光之問題發生，GMAC 針對此問題提出了一個綜合的方法來評估測驗試題題組。
4. 試題偏差：傳統調查試題偏差的方式是透過前測來偵測試題偏差，不過真正的測驗資料卻可以提供前測所無法提供的資訊。GMAC 提供了一個實務的替代觀點，來偵測題庫的試題參數對每一個類別的應考人是否合適。
5. 試題參數修正：適性化測驗的試題有非常長的保存期限，本研究對於試題參數經過若干時間後是否仍具有一致性提供一些實用的檢測方法，用來檢測試題參數長時間之一致性。

伍、心得與建議

一、 規劃時期之評估要項

一個教育測驗之主管機關在規劃是否導入電腦適性化測驗時，並非單純的是非題，而須多方衡酌並慎重考量；經由會議所觀察之各國發展適性化測驗經驗，建議未來台灣規劃測驗之主管機關在導入適性化測驗之前，可針對下列要項加以評估與分析，俾利提供應考人最佳服務。

(一)借重專家學者及委外開發

會議中多個國家在分享適性化測驗發展經驗中均提及適性化測驗之發展與建置係採分工合作之方式，針對題庫建置、適性化測驗演算法的設計及信、效度驗證多透過一個獨立的第三者（如 GMAC 之 GMAT）或教育心理學校之教育心理學者（如西班牙 eCAT 系統）來規劃與分析，而適性化測驗之建置與推展則以委外開發之方式辦理（如 GMAC 之 GMAT、西班牙 eCAT 系統、新加坡 ESS 系統）；有鑑於各國發展適性化測驗之經驗，未來若考選部欲發展適性化測驗時，建議委託一個獨立且具備教育心理背景之公正第三者來規劃適性化測驗之選題演算規則、試題重複使用率、試題校準與銜接等相關建置規範，再委託具適性化測驗開發經驗之整合服務廠商來建置適性化測驗施測工具，並透過前述公正客觀之第三者檢驗適性化測驗之信度、效度、檢測能力與目標達成率。

(二)周延評估所有可能成本及預期效益

GMAC 在分享適性化測驗發展經過中提及在「GMAT 發展電腦適性化測驗之評估過程中，基於報考之應考人數量逐年增加，但是卻越來越難取得應試座位，以及許多學校對於 GMAT 高得分應考人越來越難區分其差異性之問題，GMAC 開始思考解決 GMAT 上述問題之可行方案；而當 ETS 展示了電腦適性化測驗可適當解決上述兩個問題，並說明其潛在利益後，GMAC 董事會同意了 GMAT 電腦適性化測驗之轉換計畫，整個轉換的費用初步估算約需花費美金 400 至 700 萬；惟其成本估算並未包含發展新試題的額外費用，故在 GMAC 公布推動適性化測驗之政策性決策後，卻發現 GMAT 整個轉換的費用（包含發展新試題、整體建設等）共需近 1,170 萬美元。GMAT 適性化測驗建置完妥並開始推動之初，GMAC 面臨著適性化測驗等化之問

題，且無法達到預期效益；所幸經過 10 年之發展，並在 GMAC 換了另一個合作廠商、運用另一個評估等化的方法後，其發展適性化測驗的主要目的可成功達成，而侷限於試場座位數無法因應越來越多應考人之問題也獲得改善」。

透過 GMAC 分享發展適性化測驗之過程，在評估與規劃適性化測驗時，建議應先確認轉換之所有成本與目標，即周延地評估所有可能成本及預期效益，針對題庫建置（包含資訊系統委外與維護、預估建置題庫試題數量、命題審題費用、試題替換更新率與成本等）、電腦試場規模與建置洽借成本、施測工具委外開發、適性化測驗整體架構建設、延聘具備教育心理學者專家等各項成本皆須審慎評估；此外，針對適性化測驗預期目標，應檢驗期望目標與結果之達成比率，並在預定預算範圍內適時修正，以充分確認專案投資效益，並落實適性化測驗之執行率及建置目標。

(三)應用於何種類科與科目

在各國電腦適性化測驗的經驗分享中，其採用電腦適性化測驗之考試科目多為語言類科（如西班牙 eCAT、日本 JCAT）或非高度利害關係之考試科目（如以色列 AMIRAM），有鑑於各國推動電腦適性化測驗之考試多限制於某些科目，因此，在適性化測驗發展之前，應先評估是否應用適性化測驗於高利害風險之考試科目或類科（高利害風險之考試將提高試題安全風險）？是否需公布考畢試題（公布考畢試題將提高題庫建置成本）？並分析試題重複使用率與重複使用之間隔時間（試題重複使用率及時間間隔將影響試題安全性及建置成本）等相關問題，以增加推動適性化測驗之成功因素。

(四)是否採用單軌施測工具

透過此會議各國經驗分享發現：美軍軍職性向（ASVAB）及以色列高等教育測驗之舉辦皆採用雙軌施測工具，亦即同時採用紙本測驗與電腦化適性測驗；此兩種電腦適性化測驗皆有近 20 年的發展經驗，惟仍提供雙軌施測工具，其背景與因素值得深入探討，且應納入規劃議題加以考量；此外，一旦採用雙軌施測工具辦理考試，其兩種施測工具之試題內容規範與同等性等議題，則必須以科學公正客觀之方式加以驗證（如同以色列在發展適性化測驗時所做之兩種施測工具同等性驗證），以降低推動適性化測驗之阻力與不信任因素。

(五)以客觀量化之資料來作驗證

適性化測驗有別於傳統紙筆測驗，其在施測工具、應考人作答介面、試題呈現

方式等皆有所不同，此外，適性化測驗若依據試題反應理論（IRT）進行試題選題與考試終止之演算，其考試作答時間、考試長度（作答試題數）、試題內容更會因應考人能力值有所差異，因此，在決定採用適性化爲施測工具之前，應先以客觀量化之資料驗證電腦適性化測驗之信度與效度，並以量化之證據驗證適性化測驗之規則（如作答時間限制及變動試題數等）對每一位應考人之能力評估及考試公平性具備極高之準確度；且針對紙筆測驗與適性化測驗之同等性，更必須以實際量化資料加以驗證應考人受測表現不受施測工具不同而有所影響，以消減應考人對電腦適性化測驗之不信任，並降低其排斥行爲。

二、 建置時期之考量因素

美國研究生入學管理委員會(GMAC)在 2007 分享之 GMAT 適性化測驗建置經驗中提及：「一個成功的適性化測驗主要是對於實務問題是否有深入的研究與探討」，亦即在發展與建置適性化測驗之前，主導與管理考試之機構必須針對適性化測驗一些關鍵因素先做建置面之決定，舉凡試題的校準與銜接、施測時起始選題及適性選題之演算法、試題參數之估計與連結、能力估算之演算法、試題曝光率之控制、施測終止之規則、試題內容規範、題庫之建置等議題皆必須有所評估與選擇，以期建置最佳估算能力之適性化測驗。因此，一旦考選部欲進行適性化測驗之發展與建置時，建議應針對下列議題加以研究與確認，俾利適性化測驗之建置。

(一)分析題庫建置完整性要項與成本

一個完整的題庫可以讓適性化測驗達到以少數試題及時間測驗應考人真實能力的目的，針對題庫建置之議題，建議未來可以針對題庫試題數量與花費成本、試題與題庫之內容規範、考畢試題重複使用率與期限、試題更新頻率、試題的校準與銜接、試題參數之估算等議題多加著墨與研究，以期開發完整之題庫。

(二)規劃選題規則

1. 起始選題：適性化測驗針對應考人第一個試題須訂定選題規則（起始選題規則），如新加坡 ESS 系統第一個初始試題爲中度難度之試題，以色列電腦適性化測驗在每個領域的前兩個試題採用中難易度及低鑑別度的試題，而 GMAC 對於起始試題之挑選則提到在考試一開始時運用最大資訊演算法來挑選一個最大資訊、最具鑑別度試題的作法是很

浪費資源的，且對於鑑別應考人真實能力卻鮮少有幫助。

2. 適性選題：目前適性化測驗相關理論中針對選出最合適的下一題有許多的不同建議與看法，舉凡(1)挑選對考生能力估計提供最大訊息量的試題、(2)利用貝氏試題選擇法，或是(3)挑選難度最接近考生現階段能力估計之試題等方法皆爲常用之適性化選題方法[2]。

(三)確認能力估算演算法

目前適性化測驗相關理論中針對能力估算有許多的不同之演算法，舉凡最大概率法（Maximum Likelihood, ML）、貝式最大後驗估計法（maximum a posteriori,MAP）、貝式期望後驗估計（expected a posteriori,EAP）法等皆爲常用之能力估算演算法。

(四)定義試題曝光率及重複使用率

在適性化測驗中，試題曝光率及重複使用率對於試題安全有很大之影響，鑑於以色列及 GMAC 在推廣適性化測驗的研究報告中皆指出許多補習業者或應考人會想盡辦法拿到考過的試題，因此，定義試題重複使用率將是測驗機構在發展適性化測驗時責無旁貸的工作，以降低重複使用試題題組的風險，此外，試題曝光率則會影響投資效益及試題安全性，若選題規則導致部分試題曝光率過高，而部分試題曝光率甚低，則將無法正確成功地估算應考人真實能力，也浪費試題開發之成本，因此，在發展適性化測驗時，須以客觀量化之資料檢驗試題曝光率，並且定義最大試題重複使用率，以期提高考試安全性及降低試題建置成本。

(五)選擇考試終止規則

適性化測驗對於考試終止之方式，有採用固定題數終止、目標訊息量終止等方式，對於適性化測驗終止規則，台灣測驗主管機關在發展適性化測驗時須延聘教育心理測驗專業評估分析是否應該限制考試長度？限制長度範圍值爲何？是否需要訂定最少應考題數？是否採用固定題數或變動考試長度？若採用變動考試長度，是否產生較多的偏誤？針對這些議題，皆須加以確定，並選擇一個最合適之終止規則來建置適性化測驗，以正確評估應考人之能力值。

(六)模擬偶發事件劇本並規劃復原程序

以色列在分享其適性化測驗建置經驗時提及:「發展適性化測驗應該至少安排兩

[2] 參考余教授民寧電腦化適性測驗之介紹。

組不同工作人員預估施測及評分階段可能遇到的各種問題，包含時間分配問題、評分問題、考試各階段所需要之復原程序等，以事前充分準備因應未來可能發生之各種偶發事件」；鑑於考試穩定性及考生公平性，上述以色列經驗尤應納入必要之開發建置程序，以因應各種電腦化測驗可能發生之問題，俾利考試之順利舉辦。

三、 赴各國深入考察與實地觀摩

(一)以色列高等教育適性化測驗

此會中投稿文章相當豐富，對於適性化測驗之發展過程也有概觀式介紹，針對會議中以色列高等教育適性化測驗發展實務與推動現況，值得我們再深入瞭解與探討，例如電腦試場建置評估與建置過程、應試座位數之安排、電腦試場與題庫之整體架構與安全性、偶發事件之處理及因應方式等，皆可再實赴該國觀摩。

(二)美國商學研究生入學測驗（GMAT）

而大會主辦單位 GMAC 所分享之 GMAT 適性化測驗發展經驗，對於我國評估適性化測驗發展之可行性亦有多方幫助，舉凡其內容規範之考量因素、題庫設計與調校、成本估算需考量之因素、評選委外建置廠商需評估之要點、與 ETS、ACT 等測驗開發廠商合作經驗、Pearson/VUE 建立與管理跨國電腦試場之經驗等，皆值得我們實地考察與學習。

(三)美國美軍軍職性向測驗（ASVAB）

此外，美軍軍職性向測驗（ASVAB）每年對百萬名軍職募兵應徵者及高中學生舉辦之美軍軍職性向測驗，約有 2/3 的募兵應徵者選擇參加電腦適性化測驗（CAT-ASVAB）之施測方式，對於應考人是否可以接受軍職訓練或是判斷該名應考人可以擔任哪一類軍職工作之性向測驗有相當大的貢獻，其性向測驗與電腦適性化測驗之推展，有將近 20 年的發展及測驗管理經驗，對於台灣強調教育與評量相輔相成之教、考、訓、用等目標有許多值得仿效與學習之內容，尤其針對美軍軍職性向電腦適性化測驗（CAT-ASVAB）在性向測驗量表之建置與演進、測量之程序、題庫之發展、電腦適性化測驗之操作介面與網路管理等議題，皆可再深入赴當地學習與觀摩。

四、 定期參與國際會議

(一)每年派員參與會議掌握資訊技術與潮流

透過此次參與 GMAC 2009 適性化測驗會議,在短短兩天內學習到最新適性化測驗理論、演算法與實務研究，並汲取各國發展適性化測驗的實務經驗，對於擴展適性化測驗的視野與見解，有其事半功倍的效果，爰此，針對參與跨國學術研討會，建議可以針對相關議題，每年派員出國研習，尤其針對資訊技術快速演變之趨勢，可透由專家學者經驗交流與分享，即時獲取最新發展潮流，並且針對國內現況截長補短，在資訊業務上有所改善與創新。

(二)與國際專家學者建立聯繫管道

另針對參與國際會議之準備事項，可分爲與會前、會議中及與會後加以說明；因國際會議專家學者所發表之語言並非母語，建議應於事前與主辦單位聯繫，請主辦單位提供相關議題簡報資料與摘要，俾利事前瀏覽會議主題並於會前確認有興趣之議題；此外，會議期間除了專心學習與互動外，與國外專家友人自我介紹自當不可缺少，一來透過會議建立國際友誼，二來則可建立專家學者資料庫，當新種資訊業務構思與規劃上有所瓶頸時,自可聯繫國際友人適時提供協助與經驗分享;最後，與會後須建立聯繫與訊息交換之管道，於參加會議回國之後，建議適時地和與會友人問候、聯絡與保持日常業務聯繫，並針對相關議題與研究，詢問主講者是否協助提供投稿論文供參，經學習瞭解後，更進一步針對疑惑提出相關問題以深入探討，並轉換爲對業務有幫助之新知與想法。

附錄一：會議行程表

2009 GMAC® CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING

~ Preliminary Schedule ~

MONDAY, JUNE 1	
PRE-CONFERENCE SESSIONS (Separate registration required.)	
9:00 am—12:15 pm	**IRT for CAT** *(Including: basic assumptions of CAT, IRT models and their parameters, θ estimation and standard error, estimating item parameters, information for dichotomous and polytomous models, and linking item parameters to create an item bank)*
12:15—1:45 pm	Lunch
1:45—5:00 pm	**Essentials of CAT** *(Including: basic elements of a CAT, types of CAT, CAT versus sequential testing, what to consider before implementing CAT, implementing a live CAT, and operational issues for some CATs)*
TUESDAY, JUNE 2	
8:30—9:00 am WELCOME	**Conference Opener** *Lawrence M. Rudner, Graduate Management Admission Council®* *David J. Weiss, University of Minnesota at Twin Cities*
9:00—10:15 am	**Realities of CAT** *Effect of Early Misfit in Computerized Adaptive Testing on the Recovery of θ* Rick Guyer and David J. Weiss, University of Minnesota *Quantifying the Impact of Compromised Items in CAT* Fanmin Guo, Graduate Management Admission Council *Guess What? Score Differences with Rapid Replies versus Omissions on a Computerized Adaptive Test* Eileen Talento-Miller and Fanmin Guo, Graduate Management Admission Council *Termination Criteria in Computerized Adaptive Tests: Variable-Length CATs Are Not Biased* Ben Babcock and David J. Weiss, University of Minnesota
10:15—10:30 am	Refreshment Break
10:30 am—12:00 pm	**CAT for Classification** *Computerized Classification Testing in More Than Two Categories by Using Stochastic Curtailment* Theo J.H.M. Eggen and Jasper T. Wouda, CITO, Arnhem, The Netherlands *Utilizing the Generalized Likelihood Ratio as a Termination Criterion* Nathan A. Thompson, Assessment Systems Corporation *Adaptive Testing Using Decision Theory* Lawrence M. Rudner, Graduate Management Admission Council *"Black Box" Adaptive Testing by Mutual Information and Multiple Imputations* Anne Thissen-Roe, Kronos *A Comparison of Computerized Adaptive Testing Approaches: Real-Data Simulations of IRT- and Non-IRT-Based CAT with Personality Measures* Monica M. Rudick, Wern How Yam, and Leonard Simms, University of Buffalo
12:00—1:00 pm	Lunch

2009 GMAC® CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING

~ Preliminary Schedule ~

TUESDAY, JUNE 2

12:30—2:00 pm

POSTER SESSION

CAT Research and Applications Around the World

A Comparison of Three Methods of Item Selection for Computerized Adaptive Testing
Denise Reis Costa, Camila Akemi Karino, and Fernando A.S. Moura, University of Brasilia, Dalton F. Andrade, INE-CTC/UFSC, Brazil

Adequacy of an Item Pool for Proficiency in English Language from the University of Brasilia for Implementation of a CAT Procedure
Camila Akemi Karino, Denise Reis Costa, and Jacob Arie Laros, CESPE/University of Brasilia, Brazil

Development of an Item Model Taxonomy for Automatic Item Generation in Computerized Adaptive Testing
Hollis Lai, Mark J. Gierl, and Cecilia Alves, University of Alberta, Canada

An Approach to Implementing Adaptive Testing Using Item Response Theory in a Paper-Pencil Mode
V. Natarajan, MeritTrac Services Pvt. Ltd, India

Assessing the Equivalence of Internet-Based vs. Paper-and-Pencil Psychometric Tests
Naomi Gafni, Keren Roded, and Michael Baumer, National Institute for Testing and Evaluation, Israel

Features of a CAT System and Its Application to J-CAT
Shingo Imai et al., Yamaguchi University, Japan

Adaptive Measurement of Cognitive Ability Based on a Person's Zone of Nearest Development
Marina Chelyshkova and Victor Zvonnikov, State University of Management, Russia

Implementing Figural Matrix Items in a Computerized Adaptive Testing System: Singapore's Experience
Tay Poh Hua and Raymond Fong, Ministry of Education, Singapore

Constrained Item Selection Using a Stochastically Curtailed SPRT
Jasper T. Wouda and Theo J.H.M. Eggen, CITO, The Netherlands

Using Enhanced Effective Response Time to Detect the Extent and Track the Trend of Item Pre-Knowledge on a Large-Scale Computer Adaptive Assessment
Jie Li and Xiang Bo Wang, ACT, Inc., United States

Computerized Adaptive Testing for the Singapore Employability Skills System (ESS)
Patrick Rickard, CASAS; James B. Olsen, Alpine Testing Solutions; Debalina Ganguli, CASAS; and Richard Ackermann, Team Code, Inc., United States

Criterion-Related Validity of an Innovative CAT-Based Personality Measure
Robert J. Schneider, PDRI; Richard A. McLellan, PreVisor, Inc.; Tracy M. Kantrowitz, PreVisor, Inc.; Janis S. Houston, PDRI; Walter C. Borman, PDRI; United States

1:00—1:40 pm

CAT in Spain and Israel

Computerized Adaptive Testing in Spain: Description, Item Parameter Updating and Future Trends of eCAT
Francisco J. Abad, Universidad Autonóma de Madrid; David Aguado, Universidad Autonóma de Madrid; Juan Ramón Barrada, Universidad Autonóma de Barcelona; Julio Olea, Universidad Autonóma de Madrid; Vicente Ponsoda, Universidad Autonóma de Madrid, Spain

Twenty-Five Years of Applying CAT for Admission to Higher Education in Israel
Naomi Gafni, National Institute for Testing and Evaluation, Jerusalem, Israel

2009 GMAC® CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING

~ Preliminary Schedule ~

TUESDAY, JUNE 2	
2:00—3:15 pm CONCURRENT SESSION I	**Item Selection** *Item Selection and Hypothesis Testing for the Adaptive Measurement of Change* Matthew Finkelman, Tufts University School of Dental Medicine; David J. Weiss, University of Minnesota; and Gyenam Kim-Kang, Korea Nazaren University *A Gradual Maximum Information Ratio Approach to Item Selection in Computerized Adaptive Testing* Kyung (Chris) T. Han, Graduate Management Admission Council *Item Selection with Biased-Coin Up-and-Down Designs* Yanyan Sheng, Southern Illinois University at Carbondale *A Burdened CAT: Incorporating Response Burden with Maximum Fisher's Information for Item Selection* Richard J. Swartz, The University of Texas M.D. Anderson Cancer Center, and Seung W. Choi, Northshore University Health System Research Institute and Northwestern University
2:00—3:15 pm CONCURRENT SESSION II	**Real-Time Analysis** *Adaptive Item Calibration: A Simple Process for Estimating Item Parameters Within a Computerized Adaptive Test* G. Gage Kingsbury, Northwest Evaluation Association *On the Fly Item Calibration in Low States CAT Procedures* Sharon Klinkenberg, Department of Psychology, University of Amsterdam; Marthe Straatemeier, Department of Psychology, University of Amsterdam; Gunter Maris, CITO; and Han van der Maas, Department of Psychology, University of Amsterdam *An Automatic Online Calibration Design in Adaptive Testing* Guido Makransky, University of Twente/Master Management International A/S, and Cees A. W. Glas, University of Twente *Investigating Cheating Effects on the Conditional Sympson and Hetter Online Procedure with Freeze Control for Testlet-Based Items* Ya-Hui Su, University of California, Berkeley
3:15—3:25 pm	Refreshment Break
3:25—5:30 pm US GOVERNMENT SUPPORTED CAT PROGRAMS AND PROJECTS	**Department of Defense** *The Nine Lives of CAT-ASVAB: Innovations and Revelations* Mary Pommerich, Daniel O. Segall, and Kathleen E. Moreno, Defense Manpower Data Center **National Institutes of Health** *The CAT-DI Project: Development of a Comprehensive CAT-Based Instrument for Measuring Depression* Robert D. Gibbons, University of Illinois at Chicago *Development of a CAT to Measure Dimensions of Personality Disorder: The CAT-PD Project* Leonard J. Simms, University of Buffalo *The MEDPRO Project: An SBIR Project for a Comprehensive IRT and CAT Software System* *IRT Software* — David Thissen, Scientific Software International *CAT Software* — Nathan Thompson, Assessment Systems Corporation
7:00 pm	Reception

2009 GMAC® CONFERENCE ON COMPUTERIZED ADAPTIVE TESTING

~ Preliminary Schedule ~

WEDNESDAY, JUNE 3	
8:15—9:25 am CONCURRENT SESSION III	**Item Exposure** *Reviewing Test Overlap Rate and Item Exposure Rate as Indicators of Test Security in CATs* Juan Ramón Barrada, Julio Olea, Vicente Ponsoda, and Francisco J. Abad, Universidad Autónoma de Madrid, Spain *Optimizing Item Exposure Control and Test Termination Algorithm Pairings for Polytomous Computerized Adaptive Tests with Restricted Item Banks* Michael Chajewski and Charles Lewis, Fordham University *Limiting Item Exposure for Key-Difficulty Ranges in a High-Stakes CAT* Xin Li, Kirk A. Becker, and Jerry L. Gorham, Pearson VUE
8:15—9:25 am CONCURRENT SESSION IV	**Multidimensional CAT** *Comparison of Adaptive Bayesian Estimation and Weighted Bayesian Estimation in Multidimensional Computerized Adaptive Testing* Po-Hsi Chen, Taiwan Normal University *Comparison of Ability Estimation and Item Selection Methods in Multidimensional Computerized Adaptive Testing* Qi Diao and Mark Reckase, Michigan State University *Multidimensional Adaptive Testing: The Application of Kullback-Leibler Information* Chun Wang and Hua-Hua Chang, University of Illinois at Urbana-Champaign *Multidimensional Adaptive Personality Assessment: A Real-Data Confirmation* Alan D. Mead, Avi Fleischer, and Jessica Sergent, Illinois Institute of Technology
9:35—10:45 am	**Item and Pool Development** *A Comparison of Three Procedures for Computing Information Functions for Bayesian Scores from Computerized Adaptive Tests* Kyoko Ito, Human Resources Research Organization; Mary Pommerich, Defense Manpower Data Center; and Daniel O. Segall, Defense Manpower Data Center *Adaptive Computer-Based Tasks Under an Assessment Engineering Paradigm* Richard M. Luecht, The University of North Carolina t Greensboro *Developing Item Variants: An Empirical Study* Anne Wendt, National Council of State Boards of Nursing; Shu-chuan Kao, Pearson VUE; Jerry Gorham, Pearson VUE; and Ada Woo, National Council of State Boards of Nursing *Evaluation of a Hybrid Simulation Procedure for the Development of Computerized Adaptive Tests* Steven W. Nydick and David J. Weiss, University of Minnesota
10:45—11:00 am	Refreshment Break
11:00—11:55 am	**Diagnostic Testing** *Computerized Adaptive Testing for Cognitive Diagnosis* Ying Cheng, University of Notre Dame *Obtaining Reliable Diagnostic Information through Constrained CAT* Hua-Hua Chang, Jeff Douglas, and Chun Wang, University of Illinois *Feasibility of Applying the DINA Cognitive Diagnostic Model to Content Mastery on a Large-Scale Computerized Adaptive Assessment* Alan Huebner, Xiang Bo Wang, and Sung Lee, ACT, Inc.
11:55 am—12:30 pm	**Wrap-Up and Future Directions**

附錄二：會議議程摘要

The GMAC® 2009 Computerized Adaptive Testing Conference

An International Conference on CAT Methods and Applications

Sponsored by the Graduate Management Admission Council

ABSTRACTS

Effect of Early Misfit in Computerized Adaptive Testing on the Recovery of θ

Rick Guyer and David J. Weiss, *University of Minnesota*

This study focused on how early person misfit affected the recovery of θ for a computerized adaptive test (CAT) basd on the 3-parameter logistic model. Number of misfitting items, generating θ, item selection method, and θ estimation method were independent variables in this study. The number of misfitting initial item responses was varied from $k = 0$ to 4 items. Ten different generating θ values were used at intervals from –3 to +3. For the five conditions in which θ was less than or equal to 0, the first k responses were fixed to be correct; for the five conditions where θ was greater than or equal to 0, misfit was introduced by fixing the first k item responses to be incorrect. Maximum likelihood, weighted likelihood (WLE), and expected a posterior (EAP) estimation were used to estimate θ. Both Fisher information and Kullback-Leibler information item selection methods were used. All independent variables were crossed in the simulation design, with 1,000 simulees per cell. Recovery of θ was indexed by bias, standard error, and root-mean-square error at CAT lengths of 15, 25, 35, and 50 items. ANOVA was used to analyze the results and major effects were identified by eta-squared.

It was found that CAT could recover from misfit-as-correct-responses (MCR) for low ability simulees given a sufficient number of items. CAT could not recover from misfit-as-incorrect-responses (MIR) for high ability simulees, even after 50 items. At 50 items, a small amount of bias was observed for 1 misfitting item; as the number of misfitting items increased to 4, the bias increased and was substantial for all positive values of θ. The differences between the Fisher and Kullback-Leibler information-based item selection dissipated after 15 items were administered – with one exception: for the MIR conditions, it was found that WLE functioned differently under the two item selection methods even after 50 items were administered. A follow-up study was performed, and it was found that WLE was highly sensitive to item difficulty early in the CAT. Implications of the results and suggestions for future research will be provided.

For further information: guyerr@assess.com or guyer005@umn.edu

Quantifying the Impact of Compromised Items in CAT

Fanmin Guo, *Graduate Management Admission Council*

If a few test items should become compromised, their impact on test scores would not be constant across different computerized adaptive test (CAT) programs. For the same number of items compromised, the impact might be more serious in some CAT programs than others because the impact interacts with the complexity of the test specification, size of CAT pools, item exposure control, item selection algorithm, and scoring method employed in CAT programs. As a result, evaluating the impact of compromised items on test scores in a CAT program is not easy. Most of the previous research focused on the impact on a group of examinees through simulations.

In this study, a new method of simulation is introduced that focuses on the impact on individual examinees using the GMAT® CAT as an example. For each simulee, two paths of simulations were run. The first path is the conventional simulation under no compromised item condition. The second path follows the selected items and response patterns in the first path until a "compromised" item is "administered." Then the answer to this item is reset to a correct answer to simulate a "security breach." After that, the path branches to selecting new items. All the answers to subsequent "compromised" items are set as correct answers until the end of the test. The purpose of this method is to quantify the impact of compromised items as well as its interaction with the item selection and other CAT operational configurations. Since each simulee will have two scores from the two separate paths, this method allows estimating the range of score gains and the number of compromised items seen by each individual. It allows reports that, if n items from a CAT pool were exposed to m examinees, x examinees would gain y score points due to the impact of compromised items. The method employed in this study applies to any CAT program.

For further information: fguo@gmac.com

Guess What? Score Differences With Rapid Replies Versus Omissions On A Computerized Adaptive Test

Eileen Talento-Miller and Fanmin Guo, *Graduate Management Admission Council*

Estimation of ability in computerized adaptive testing relies on the assumption that examinees are responding based on their content knowledge and skills. Guessing might have differential consequences on scoring depending on the situation. In the case of time constraints, examinees are faced with a choice of leaving questions blank or randomly responding. The current study provides guidance for examinees based on real data from an operational CAT. Previous research provides an incomplete picture of the effects of choosing a guessing strategy versus omitting items in the scoring of an operational CAT. The study expands on previous research by using operational as opposed to simulated data, comparing results in verbal and quantitative sections of a test, and framing the results to provide guidance for examinees.

In this study, scores from tests with responses classified as random guesses are compared to scores that would be observed if the items had not been reached. Items are classified as guesses by examining the distribution of latency for correct responses to determine a rapid guessing threshold. The threshold is

then checked against the proportion of correct responses at that level to find close to chance levels of correctly answering the item. The guessing threshold of 10 seconds for verbal items and 7 seconds for quantitative items was applied to all item positions. Only consecutive rapid guesses at the end of the section were examined. Scores of examinees who guessed are recalculated to reflect ending the test and omitting the remaining items rather than guessing. Although the results tend to favor guessing as a strategy, the degree of difference varied based on section content, number of items involved, and estimated ability of the examinee. In the verbal section of the test, few differences existed between guessing scores and omit scores. In the quantitative section, the benefit of guessing became more pronounced as the number of items increased. The results are particularly intriguing when ability groups are compared. Both the verbal and quantitative sections show a slight preference for the omit strategy in the low ability group. For the high ability group, apparently severe penalties for omissions in the shorter quantitative measure appear to make guessing the unequivocal strategy of choice. Future research could include more definitive methods for determining random guessing and examine guessing at different positions within the test rather than merely at the end. Ultimately, the advice for candidates remains the same for a CAT as it would for other tests: Time management is important to allow ample opportunity to give thought to every question.

For further information: talento-miller@gmac.com

Termination Criteria in Computerized Adaptive Tests: Variable-Length CATs Are Not Biased

Ben Babcock and David J. Weiss, *University of Minnesota*

This simulation study examined the performance of several CAT termination rules: four basic termination rules (standard error, minimum information, change in θ, and fixed length) and two combinations of standard error and minimum information termination. Four item banks were used: a flat information bank with 500 items, a peaked information bank with 500 items, a flat information bank with 100 items, and a peaked information bank with 100 items. Maximum likelihood scoring was used to estimate θ. For non-mixed response vectors, θ was incremented by 0.5. In addition to examining the performance of these termination criteria, the study was concerned with further examining the conclusion from previous research that variable-length CATs are more biased than fixed-length CATs (Chang & Ansley, 2003; Yi, Wang, & Ban, 2001).First, a number of variable-length CAT conditions were simulated. Then, the mean number of items administered for selected variable-length conditions was determined and fixed-length CATs were simulated with the appropriate number of items in order to properly compare variable- and fixed-length CATs. CAT performance was compared in terms of test length, as well as bias, RMSE, and correlation in the recovery of true θ.

As expected, longer CATs yielded more accurate θ estimation no matter which termination criterion was used, but there were diminishing returns with a large numbers of items. It is recommended that CATs should administer a minimum number of 15 to 20 items to ensure stable measurement. The standard error termination rule, also known as the equiprecise measurement rule, performed the best among all the methods if the standard error cutoff was sufficiently low and the item bank contained the

amount of Fisher information needed to reach the cutoff. Standard error termination was also quite efficient by administering relatively few items. Change in θ, a newer termination criterion, performed slightly worse than its fixed-length termination counterpart. Hybrid termination rules, such as combining minimum information and standard error termination, functioned the best when the item bank was small but had a peaked information function. The fixed-length CATs did not perform better than their standard error termination counterpart when equated for average test length. Previous findings stating that variable-length CATs are more biased than fixed length CATs were the result of two procedural artifacts in prior research: (1) variable-length CATs were generally much shorter than the fixed-length CATs; and (2) most previous studies used Bayesian scoring, which biased the shorter variable-length CATs in the previous studies because the prior has more of an effect on θ estimation when there is less psychometric information. Standard error termination actually performed slightly better than fixed-length CATs of comparable mean length in estimating low true θ values.

For further information: babco062@umn.edu

Computerized Classification Testing in More Than Two Categories by Using Stochastic Curtailment

Theo J. H. M. Eggen, *CITO and University of Twente, The Netherlands*
Jasper T. Wouda *CITO, The Netherlands*

When classification into a limited number of categories is the main purpose of testing, algorithms based on the application of sequential statistical testing have shown to be better performing alternatives above traditional estimation based computerized adaptive tests(e.g. Reckase, 1983 and Eggen & Straetmans, 2000) In these studies, the sequential probability ratio test (SPRT; Wald, 1947) is applied in order to decide whether more observations on items are needed and which classification decision is to be made. When a decision cannot be made with the predetermined decision error rates, in practice the procedure is always truncated at a maximum test length. Recently Finkelman (2003, 2008) proposed an adaptation of stochastic curtailment with which he created an additional stopping rule for the SPRT. This "stochastically curtailed sequential probability ratio test:, or SCSPRT, generally follows the same rules as the conventionally truncated SPRT, including its stopping rule. However, the SCSPRT adds some rules in order to be able to stop testing in the cases where a change in decision between categories is possible, but unlikely. Finkelman (2003) introduced the method for the case of classifications in two categories and items selected to be most informative at the classification point. In this paper the generalization of the application of the SCSPRT to problems with more than two categories is discussed with a focus on the problems encountered in generalizing to the three-category problem. In general the (optimal) composition of the test cannot be fixed in advanced when there is more than one cutting point, which is a requirement of Finkelman's SCSPRT. The way the application of stochastic curtailment in combinations of SPRTs can be combined with the adaptive item selection in the test is described. The performance of the proposed procedures is illustrated by results of simulation studies.

For further information: Theo.Eggen@cito.nl

Utilizing the Generalized Likelihood Ratio as a Termination Criterion

Nathan A. Thompson, *Assessment Systems Corporation*

A common application for adaptive testing is to classify examinees into mutually exclusive groups. Currently, the predominant psychometric termination criterion for designing computerized classification tests is the sequential probability ratio test (SPRT; Reckase, 1983) based on item response theory. This operates by formulating a hypothesis test that a given examinee's ability value θ is equal to a fixed value below (θ_1) or above (θ_2) the classification cutscore. Recently, it was demonstrated that the SPRT, which only uses fixed values, is less efficient than a generalized form which tests whether a given examinee's θ is *below* θ_1 *or above* θ_2 (Thompson, 2007). Moreover, this better represents the conceptual purpose of the exam, which is to test whether θ is above or below the cutscore.

The purpose of this study was to explore the specifications of the new generalized likelihood ratio (GLR). As with the SPRT, the efficiency of the procedure depends on the nominal error rates and the distance between θ_1 and θ_2 (Eggen, 1999). Preliminary results suggest that observed error rates are closest to nominally specified error rates when the values of θ_1 and θ_2 are approximately 0.1 from the cutscore. The study utilized a monte carlo approach, with 10,000 examinees simulated under each condition. Three levels of nominal accuracy were investigated (90%, 95%, and 99%), as well as 25 values of the difference between the cutscore and θ_1 or θ_2 (0.00 to 0.50 in increments of 0.2). Additionally, another formulation was investigated that forms the likelihood ratio based on an integration of the likelihood function. This was also suggested by Thompson (2007), but was not accurate due to the asymmetry of the likelihood function when the three-parameter model is used; the left-hand end of the likelihood function is substantially higher than the right-hand end because of the *c* parameter. This artificially biases the ratio in the negative direction. Methods of correcting for this are suggested.

For further information: nthompson@assess.com

Adaptive Testing Using Decision Theory

Lawrence M. Rudner, *Graduate Management Admission Council*

In the introduction to their classic textbook, Cronbach and Gleser (1957) argue that the ultimate purpose for testing is to arrive at classification decisions. Many of today's decisions are indeed binary, e.g., whether to hire someone, whether a person has mastered a particular set of skills, or whether to certify an individual. Categorical, as opposed to continuous, outcomes are also common, e.g., the percent of students that perform at the basic, proficient, or advanced level in state assessments. IRT models have been applied to help make classification decisions by laboriously placing individuals on ability scales and then using cut-points to make classifications. IRT models, however, are not always applicable in practical situations. IRT is fairly complex, relies on several fairly restrictive assumptions, requires large calibration samples, and might not make efficient use of test questions when the goal is simple classification. This paper presents an alternative underlying model for adaptive testing using measurement decision theory and then compares those procedures with IRT in terms of classification accuracy using two sets of simulated item response data. The research examines three ways to

adaptively select items using decision theory: a traditional decision theory sequential testing approach (expected minimum cost), information gain (modeled after Kullback-Leibler, 1951), and maximum discrimination. It also examines the use of Wald's (1947) well-known sequential probability ratio test (SPRT) as a test termination rule in this context.

Initial results show that the minimum cost approach was notably better than the best-case possibility for IRT. Information gain, which is based on entropy and comes from information theory, was almost identical to minimum cost. The simple approach using the item that best discriminates between the two most likely classifications also fared better than IRT, but not as well as information gain or minimum cost. Initial results also show that with Wald's SPRT, large percentages of examinees can be accurately classified with very few items. With only 25 sequentially selected items, for example, some 90% of the simulated state-NAEP examinees were classified with 86% accuracy. This is clearly a simple yet powerful and widely applicable model. The advantages of this model are many—the model yields accurate mastery state classifications, can incorporate a small item pool, is simple to implement, requires little pre-testing, is applicable to criterion-referenced tests, can be used in diagnostic testing, can be adapted to yield classifications on multiple skills, and should be easy to explain to non-statisticians. It is the author's hope that this research will capture the imagination of the research and applied measurement communities. The author can envision wider use of the model as the routing mechanism for intelligent tutoring systems. Items could be piloted with a small number of examinees to vastly improve end-of-unit examinations. Certification examinations could be created for specialized occupations with a limited number of practitioners available for item calibration. Short tests could be prepared for teachers to help make tentative placement and advancement decisions. A small collection of items from one test, say state-NAEP, could be embedded in another test, say a state assessment, to yield meaningful cross-regional information.

For further information: LRudner@gmac.com

"Black-Box" Adaptive Testing by Mutual Information and Multiple Imputations

Anne Thissen-Roe, *Kronos*

Over the years, most CAT systems have used score estimation procedures from item response theory. IRT models have salutary properties for score estimation, error reporting, and next-item selection. However, some testing purposes favor scoring approaches outside IRT. Where a criterion metric is readily available and more relevant than the assessed construct, for example in the selection of job applicants, a predictive model might be appropriate (Scarborough & Somers, 2006). Neither IRT scoring nor unidimensional assessment structure can be assumed. Yet, the primary benefit of CAT remains desirable: shorter assessments with minimal loss of accuracy due to unasked items. Without IRT, it remains possible to create a CAT system that produces an estimated score from a subset of available items, recognizes differential item information given the emerging item response pattern, and optimizes the accuracy of the score estimated at every successive item. No information is needed about the internal mechanisms of the scoring algorithm, provided it has certain properties: (1) The score must be discrete or able to be made discrete, such as by application of cut scores or reporting of integer scale scores. The score can be a nominal category; and (2) The degree to which the score

changes when a particular item response is given must vary based on the responses to other items. If these conditions are met, the scoring algorithm can be treated as a "black box," with adaptation conducted on the outside. The method of multiple imputations (Rubin, 1987) might be used to simulate plausible scores given plausible response patterns to unasked items (Thissen-Roe, 2005). This method is also capable of rendering an estimate of the error introduced by unasked questions. Mutual information might then be calculated in order to select an optimally informative next item (or set of items). This is related but not identical to the methods of Weissman (2007) for item selection, and Chambless and Scarborough (2001) for feature selection.

Two neural network-centered scoring algorithms serve as structural examples. In early testing, previously observed response patterns to the complete assessments were resampled according to CAT item selection. The reproduced CAT scores were compared to full-length assessment scores. Approximately 95% accurate assignment of examinees to one of three score categories was achieved with a 70%-80% reduction in median test length. This method of CAT is more computationally demanding than traditional IRT-based approaches, due to the necessity of completely scoring some hundreds or thousands of response patterns per item selected. Factors influencing performance were also examined during early testing. Reducing the number of multiple imputations used is a way of reducing computation time; it appears to impact assignment accuracy less than limiting items presented under a confidence-based stopping rule. Computation time can also be reduced by sacrificing algorithmic simplicity to move repeated computations outside of the "black box;" however, such shortcuts impose a maintenance burden. Mixing "black box" CAT with Internet testing also requires minimizing the data size and frequency of transactions between client and server, for which the simplest algorithm is well suited.

For further information: anne.thissenroe@kronos.com

A Comparison of Computerized Adaptive Testing Approaches: Real-data Simulations of IRT- and Non-IRT-based CAT with Personality Measures

Monica M Rudick, Wern How Yam, and Leonard Simms
University at Buffalo, State University of New York

A variety of approaches have been implemented to create CAT personality assessments. Recent research has focused on IRT for CAT personality measures, although its use is both computationally complex and requires certain assumptions to be met that do not always hold for personality measures. As a result, non-IRT-based CAT approaches, such as the countdown method, have also successfully been applied to CAT versions of personality measures. In the countdown method, there is some debate regarding whether classification or full-scores-on-elevated-scales (FSES) methods are more preferable. In addition, it is unclear how order of item administration might impact item savings and the validity of scores. Both IRT and non-IRT based methods appear to yield numerous advantages for CAT assessments, most notably time and item savings, and ease of administration. However, these two methods have yet to be directly compared. The purpose of the present study was to compare non-IRT and IRT-based approaches utilizing real-data CAT simulations on a large diverse sample (N = 8,690) who completed the Schedule for Nonadaptive and Adaptive Personality (SNAP). The report focuses

on the three longest SNAP Scales: Disinhibition (DIS), Negative Temperament (NT) and Positive Temperament (PT). Simulation analyses compared item savings, item and test information, test validity, and fidelity across the IRT- and non-IRT CAT methods. In addition, within the countdown method simulations, the simulations examined whether item presentation order impacted the results. Results will have implications for test developers wishing to apply CAT technology to personality measures.

For further information: mmrudick@buffalo.edu

A Comparison of Three Methods of Item Selection for Computerized Adaptive Testing

Denise Reis Costa, Camila Akemi Karino, *CESPE/University of Brasilia,* Brazil
Fernando A. S. Moura, *Federal University of Rio de Janeiro,* Brazil
Dalton F. Andrade, *Federal University of Santa Catarina,* Brazil

One of the most important components of CAT is the set of procedures for item selection. Unlike traditional paper-and-pencil tests, adaptive procedures administer items that fit the examinee's level of proficiency. This selection is based both on the characteristics of the items (e.g., item difficulty or discrimination parameters) and on the estimated proficiency of the examinee. This study is a work-in-progress that aims to evaluate the performance of three different CAT item selection methods: the first one is derived from the maximum information criterion, one of the most popular item selection methods inCAT; the second method is based on the global information method as defined by Chang and Ying (1996), which use the Kullback-Leibler measure, while the third selection method based on the predictive analysis defined by the expected maximum information criterion proposed by van der Linden (1998). To evaluate the three different methods, the answers of ten examinees with different skill levels were simulated for an item pool containing 246 items of the Instrumental English test of the University of Brasilia. The resulting database was fit by a three-parameter logistic model on a scale with mean 0.0 and standard deviation of 1.0, later transformed into a mean of 100 and standard deviation of 25. The examinees' iterative proficiencies were estimated using expected a posteriori (EAP). An initial analysis of bias and mean square error suggested that all methods performed similarly to estimate examinees' proficiency. However, databank-related characteristics might have influenced those measures, since it is not yet an ideal item pool for CAT implementation. With these results, it can be concluded that there is no apparent statistical difference in relation to the proficiency estimation for the three presented methods for the analyzed item bank.

For further information: denise@cespe.unb.br

Adequacy of an Item Pool For Proficiency in English Language From The University of Brasília For Implementation of a CAT Procedure

Camila Akemi Karino, Denise Reis Costa, and Jacob Arie Laros
CESPE/University of Brasilia, Brazil

The possibility of applying different item sets according to the level of ability of each respondent has stimulated, among other factors, an increasing use of CAT. In spite of the increasing use, this study is

one of the first initiatives in this field in Brazil. The item pool used in this study is a database of the proficiency exam in English language has been in use since 2004 by the University of Brasilia. This exam aims to assess the student's comprehension of texts in the English language. The exam is a paper-and-pencil test that is composed of 50 multiple-choice items. The psychometric item quality was verified using classical test theory and IRT. The complete item pool consists of 450 items divided into nine test forms. Each test form was responded by, on average, by 330 students. The total number of respondents was 2,969. First, each test was analyzed individually and in a second stage the nine tests were calibrated jointly. Of the 450 items, 37 items were common items between test forms. In the individual analyses, 46 items with biserial correlation less than than .20 and 80 items with discrimination parameter in the normal IRT metric less than .50 were eliminated. In the joint analysis, another 58 items with an *a* parameter less than .50 were eliminated. After the elimination of these items, the joint IRT analysis revealed a mean discrimination parameter of .77 (SD = .20), varying between .49 and 1.67. In relation to the *b* parameter, the existence of a substantial variation in difficulty level of the items was observed (varying between -3.56 and 3.23): however, the majority (75%) of the items showed a *b* parameter below .10. The median value of parameter *c* was .11 (SD =.04) with a range from .03 to .24. After the joint calibration, successive points of the scale were fixed for anchor items and each of these levels was interpreted pedagogically by specialists. The suitability of the item pool for implementation of a CAT procedure was questioned taking into consideration that 44% of the items needed to be eliminated in order to agree with pre-established psychometric criteria. Nonetheless, both the analysis of the item pool and the scale interpretation permit initial studies for the implementation of a CAT procedure. The item pool as well as the scale could be improved by repeated applications of the English exam using a CAT procedure.

For further information: camilaakarino@gmail.com

Development of an Item Model Taxonomy for Automatic Item Generation in Computerized Adaptive Testing

Hollis Lai, Mark J. Gierl, and Cecilia Alves, U*niversity of Alberta*, Canada

CAT makes tremendous demands on item banks because CATs require large numbers of test items. CATs require these item volumes for three general reasons. First, as test length increases in fixed-length CATs, requirements for test items increase to ensure that test scores are reliable (Wainer & Eignor, 2000). Second, with the emergence of cognitive adaptive tests (e.g., Zhou, Gierl & Cui, 2008), many more skills are measured at a finer grain size. Thus, more test items are required to measure these large numbers of specific skills. Third, item exposure and security concerns demand that item re-use rates be relatively small. That is, CAT requires a large number of unique test items in operational testing situations. One solution that could be developed to address these three issues is to generate many more items. Automatic item generation is an approach to item development where large numbers of offspring items (also called item instances) are generated from a parent item model. Although automatic item generation can potentially create hundreds and even thousands of items, its effectiveness is reliant on the availability of an efficient framework for creating the parent item models. The components in a parent item model for a multiple-choice item consist of the stem (the component

of an item that forms the context of the question the examinee is required to answer), the options (a set of alternatives with one correct option and multiple distracters to answer the question), and any auxiliary information (e.g., pictures, graphs).

To identify possible item model types, Gierl, Zhou, and Alves (2008) developed a taxonomy to categorize and delineate the levels of variation in components of the parent item model. One limitation of the study by Gierl et al., however, was that it focused only on mathematics items. To be applied in diverse testing situations, item models need to be created in many different content areas to allow for automatic item generation. The present study will apply the taxonomy to item models from diverse content areas, including Language Arts, Social Studies, and Science, to generate items for a computer-based testing program. While there might have been other implementations of item generation, few have been documented (Irvine, 2002). Hence, the implication of the present study is to demonstrate a systematic way to generate test items that creates large numbers of items in diverse content areas, thereby lowering the cost of item development while maintaining a high level of quality in the development process.

For further information: hollis.lai@ualberta.ca

An Approach to Implementing Adaptive Testing Using Item Response Theory in a Paper-Pencil Mode

V. Natarajan, *MeritTrac Services Pvt. Ltd,* INDIA

In India, as most of the large scale testing is conducted in the paper-pencil (offline) mode, it is important to arrive at models of implementing IRT in an offline/paper-pencil mode. MeritTrac has experimented in conducting an IRT-based test in a paper-pencil mode for the analytical abilities test for engineering graduates. With the help of item characteristics calculated prior to the test, a 6-item test with increasing item difficulty was created as a test form on paper. Normally, research shows that a 6/10 item test can be compared to 25 or more items in the test. The test was then administered to the candidates in an offline mode. The responses of the examinee were then entered in student tracking software that had been specially coded for this purpose. The output of this gives an estimation of the examinee's true score as if he/she has taken the parent 25-item test. Since it is not very feasible to conduct an online test everywhere, especially in a country like India, the importance of adaptive testing in offline mode increases many fold. In this model, we only need a single computer with student tracking software and pre-published test forms consisting of items whose characteristics have been calculated on the basis of past responses. Thus the offline mode is much more practical and is as accurate as the online mode.

In the analytical abilities test, we have looked at 100 items and the responses of 1,000+ examinees on each of these items, which we entered into BILOG and item difficulty values were generated. 93 items were found to be relevant and the parent test of 93 items eventually emerged. The items were grouped into 6 groups and 10 items were selected (one item each very easy and easy two items from below average, average difficult and very difficult). Several sets of 10-item adaptive tests each were selected and administered to the examinees. Their responses to 10 items were categorized in terms of 9,8,7,6,5,4,3,2,1 correct and a table generated from which ability and true scores can be read. In this

methodology, the test administrator needs to be very cautious when dealing with student tracking software so that mistakes are not made in entering the values of item numbers in the reshuffled version and the examinee's responses.

For further information: madan@merittrac.com

Assessing the Equivalence of Internet-Based vs. Paper-and-Pencil Psychometric Tests

Naomi Gafni, Keren Roded, and Michal Baumer
National Institute for Testing and Evaluation, Israel

Few studies have yielded information regarding the equivalence of high-stakes admissions tests administered via the Internet and paper-and-pencil administrations of those tests (Potosky & Bobko, 2004). Despite the lack of evidence regarding the equivalence of scores obtained in these two modalities, there is increasing demand for Internet-based testing, with the number of recruitment and admissions tests administered via the Internet constantly rising. This is largely due to the convenience and efficiency that the medium offers. The Psychometric Test, which is used for admission to institutions of higher education in Israel, is a high-stakes examination. The test consists of three sections: Verbal Reasoning (60 items), Quantitative Reasoning (60 items), and English as a Foreign Language (54 items). All items are in multiple-choice format. At the present time, most of the examinees take the paper-and-pencil version of the test. It is anticipated that Internet-based administration will be expanded. Given that this process will be gradual, and for a period of time the test will be administered in two parallel modalities, establishing the equivalence of scores is of paramount importance.

The goal of the present study was to compare the achievement of examinees who took the paper-and-pencil version of the Psychometric Test with the achievement of those who took it via the Internet. The question of equivalence arises because there are certain differences between a linear computerized test and a traditional paper-and-pencil test, and also between computerized tests administered via the Internet and those that are not. In the former case, the differences lie in the presentation of the items, the method of answering, how reading comprehension passages and questions with graphic components are presented, and in how time is allotted. Internet-based administration brings other factors into play, for example interruptions to the power supply, non-standard computers in different laboratories, Internet server problems, the impact of heavy traffic on the server, a greater risk of items being compromised and the challenge of handling problems during the administration itself. The relationship between performance on the experimental test and several background variables (based on a feedback questionnaire) was also examined. The participants were 381 examinees who registered for the October 2008 administration of the Psychometric Test. The paper-and-pencil version was given to 192 of these participants, and 189 were tested via the Internet. Assignment to the two groups was random. 370 of the participants in the experiment (185 from each one of the groups) took the actual Psychometric Test a month after the experimental administration.

The following conclusions are based on analysis of the results: (1) No significant difference was found between the scores of the two groups; (2) No significant differences were found between the scores on the Verbal Reasoning and Quantitative Reasoning sections, however, the English scores were significantly higher in the computerized version, across all item types; (3). The correlation between the overall experimental scores and scores on the actual test were 0.93 and 0.94 for the computer-based and paper-and-pencil groups respectively; (4) The difference between the two groups in improvement in scores (between the experiment and actual test), both overall and for each section, was not significant; (5) The difference in scores between men and women was the same for both groups; and (6) The correlation between frequency of computer use and performance on the test was similar for both groups. Thus, it was found that the modality of administration, Internet-based or paper-and-pencil, did not affect examinee performance on the Psychometric Test. This holds with respect to item types that we suspected would become more difficult when administered by computer. The results support simultaneous administration in two modalities.

For further information: naomi@nite.org.il

Features of a CAT System and Its Application to J-CAT

Shingo Imai, Y. Akagi, *Yamaguchi University,* Japan
K. Kikuchi, *Toho University,* S. Ito, *TUFS,* Japan
Y. Nakamura, *Tokiwa University,* Japan
H. Nakasono, *Shimane University,* Japan
A. Honda, *APU,* and T. Hiramura, *TIT,* Japan

A CAT system called J-CAT or Japanese computerized adaptive test, which is operational on the internet or by LAN, has been developed and used as a proficiency test of Japanese at the college level for international students in Japan. We discuss some features of this CAT system, focused on the viewpoint of test administrators. The features discussed in this presentation include registration method, item-pool management, and utilization of test results. We illustrate how this system registers examinees and authenticates them. We also discuss how to manage an item pool; such as uploading items, setting IRT parameters, and setting answering time limits for each item. The system provides useful information for analyzing the results of a test. We highlight some features of a downloadable CSV file of properties of examinees and test results. We show what information is available for an administrator and how an administrator might utilize the information. Examinees are also provided feedback of their test results as a report form which is automatically produced at the end of a test. The system of J-CAT, which contains items for Japanese proficiency at present, can be also used for tests other than Japanese language if the items are replaced with items of other tests. The system supports Rasch, two-parameter, and three-parameter IRT models.

For further information: imai2002@yamaguchi-u.ac.jp

Adaptive Measurement of Cognitive Ability
Based on a Person's Zone of Nearest Development.

Marina Chelyshkova and Victor Zvonnikov, *State University of Management*, Russia

At the present moment the majority schools and universities of Russia attach great importance to cognitive process in education. We think that in modern testing it is important not only to estimate the degree of knowledge that the person has but also to evaluate cognitive ability, which is more complex than knowledge and skills. The measurement of cognitive ability usually requires the special content of items, which cognitive learning theories provide. But there are other aspects of such measurement. They are connected with optimization of an item's difficulty and require the application of adaptive testing .Weiss analyzed person characteristic curves and suggested some methods for adapting the test item's difficulty to the individual. These ideas were combined with the concepts of Russian scientist L. S. Vigotsky who suggested the ratio of ability to knowing something (actual zone) and ability to develop of a person's internal mental forces. His concept allows to connect the score of actual knowledge with the width of a person' s zone of the nearest development. We suggested the method for evaluating this connection by using one-parameter and two-parameter models of IRT expressed it in the form of the system of inequalities which related the person parameter and item's parameters. As applied to measurement of a cognitive ability we suggested to choose items that have difficulty appropriate to person' s zone of the nearest development instead of traditional scoring approaches in adaptive testing. We developed the connection between the width of the nearest development zone and scores of test items in terms of the difficulty and slope of item characteristic curves. It has allowed us to evaluate a person's cognitive ability and to predict his/her changes of achievement depending on the time factor and the steepness of his/her person characteristic curve. Thus, in such a way we can optimize the difficulty of test items in adaptive testing for measurement of cognitive ability.

For further information: mchelyshkova@mail.ru

Implementing Figural Matrix Items In a Computerized Adaptive Testing System – Singapore's Experience

Poh Hua Tay and Raymond Fong, *Ministry of Education,* Singapore

Figural matrix items such as Raven's Standard Progressive Matrices (SPM) are widely used for assessing general intelligence of pupils. Substantial manpower resources are incurred when administering tests on a large scale basis via paper-and-pencil (P&P). A computer-based test (CBT) would offer the advantages of logistical ease during the data collection stage, and administrative ease during the data entry stage; this is especially so for CAT, as it reduces administration time, as well. Unlike P&P and CBT, the most appropriate set of items in a CAT can be adaptively selected for each pupil based on his/her responses to previous items. This permits each pupil to be evaluated on a smaller subset of the total item pool, having better test experience as items are chosen based on his/her ability; and allows the test developer to control the error of measurement to a desired degree of precision.

In this study, an item bank of 195 figural matrix items that are similar to SPM's was created. The psychometric properties of these items were then established after trialing them on a sample of 6,821 Primary 2 pupils (equivalent to Grade 2 pupils who are about 8 years in age) of varying academic abilities from 20 coeducational schools in Singapore. IRT was used to calibrate all the figural matrix

items. From this item bank, a P&P prototype, two CAT prototypes (one starts with an easy item, while the other starts with an average item), and a CBT prototype were generated and administered, via the FastTEST Pro v2.3 platform, to four groups of Primary 2 pupils in Singapore. These groups consisted of a total of 948 Primary 2 pupils of varying academic abilities and were selected from 12 coeducational schools. SPM was also administered to all of them via P&P. This project was designed to study the comparability of the abilities of pupils estimated from the differentpPrototypes (P&P, CATs, CBT) and SPM.

For further information: tay_poh_hua@moe.gov.sg

Constrained Item Selection Using a Stochastically Curtailed SPRT

Jasper T. Wouda and Theo J. H. M. Eggen, *CITO,* The Netherlands

Computerized classification testing (CCT) can be used to increase efficiency in educational measurement. The truncated sequential probability ratio test (TSPRT) has been widely studied as a decision algorithm in CCT for two or more categories (Spray, 1993; Eggen, 1999). Finkelman (2003) added an algorithm to the TSPRT in the form of stochastic curtailment, to classify an examinee in an even earlier stage of testing. This stochastically curtailed SPRT (SCSPRT) halts testing when a change of classification is possible but unlikely. As can be seen in Finkelman (2003, 2008), the SCSPRT is an extension of the SPRT. It adds stochastic curtailment in the form of two extra stopping rules per level. Stochastic curtailment ceases testing and rejects hypothesis H_{01} if given k observations, the probability that a decision D will accept H_{01}, $Pr(D= H_{01})$, is not higher than a set value $1-\gamma$. It stops testing and accepts H_{01} if this probability is at least γ. This method makes use of the sub-optimality of the SPRT as used in truncated tests.

In the comparison of performance between the SPRT and SCSPRT (Finkelman, 2003, 2008), results showed a substantial decrease in number of items used per simulee for the SCSPRT, while the percentage of correctly classified simulees remained the same. When using real item parameters and realistic data (Wouda, 2008), this decrease became somewhat smaller, but was still substantial. However, in order to be applied in real-world tests, non-statistical constraints must also be considered. Different constraints include, for example, content balancing, answer key balancing, conflicting items and item exposure control. In this study, different constraint handling methods will be compared, together with different item selection methods. The applied constraints are content balancing and exposure control. The compared item selection methods will be selection of items at the θ estimate and selection of items at the cut-score. The methods for exposure control that will be compared for the SPRT and SCSPRT are the Sympson-Hetter method, the progressive method, and alpha-stratified testing. The methods for content balancing that will be compared are the Kingsbury and Zara (1989, 1991) approach and the weighted deviation method (WDM) by Stocking and Swanson (1993).

For further information: Jasper.Wouda@cito.nl

Using Enhanced Effective Response Time to Detect the Extent and Track the Trend of Item Pre-Knowledge on a Large-Scale Computerized Adaptive Assessment

Jie Li and Xiang Bo Wang, *ACT, Inc*.

In addition to being highly efficient and accurate in terms of scoring, diagnosis, and reporting, CAT is also known for its global ease and reach of test delivery (Wainer et al, 2000; Meijer & Nering, 1999; Parshall, Spray, Kalohn, & Davey, 2002). However, the latter advantage of CAT also introduces a tenacious problem of potentially exposing items to a high number of examinees due to its high frequency of test administration, which is likely to increase advance or pre-knowledge of items and to jeopardize score validity. Of great concern and interest to the entire educational testing industry is the possibility of validly detecting and tracking the extent that CAT items are exposed. The purpose of this research was (1) to establish population item response times for all items and associated trends for all items with a large-scale international CAT assessment and (2) to investigate the feasibility of applying "effective response time" (ERT; Meijer & Sotaridona, 2006) to detect the extent and track the trend of item pre-knowledge on suspected compromised items on this assessment. The study was based on both operational and simulated data of a large item pool of a large-scale international CAT assessment. This item pool was selected because (1) it had a substantial number of new items that were pretested in several years ago when little or no item pre-knowledge could be assumed and (2) these pretest items had a long history of operational use in subsequent years when item pre-knowledge could have been accumulated. ERT indices for both items and examinee, as described by Meijer & Sotaridona (2006), were computed against a large collection of new items at their pretest time after they passed stringent pretest item quality reviews. The ERT indices from this round were used as null hypothesis benchmarks since no serious item pre-knowledge could be assumed. In addition, simulations were conducted to project the values of these ERT indices, if examinees' response times were reduced by one-half and one-fourth, respectively. Examinees ability estimates on the operational items of this item pool were used for ERT modeling. ERT indices were also computed when all the new items were first used operationally and the results were compared with their pretest counterparts.

For further information: Jie.Li@Act.org

Computerized Adaptive Testing for the Singapore Employability Skills System (ESS)

Patricia Rickard, *CASAS,*
James B. Olsen, *Alpine Testing Solutions,*
Debalina Ganguli, *CASAS,*
and Richard Ackermann, *Team Code, Inc.*

This paper presents and demonstrates innovations in computerized adaptive testing of adult workplace literacy and numeracy skills developed by CASAS and customized for the Singapore Employability Skills System (ESS). The Singapore Workforce Development Agency (WDA) plays a pivotal role in the implementation of the ESS "to enhance the employability and competitiveness of employees and job seekers, thereby building a workforce that meets the changing needs of Singapore's economy." CASAS has designed and developed CATs for mathematics, reading, and listening, and computer-delivered tests for writing and speaking, suitable for adults. The CATs are administered in secure proctored locations using local area networks and an electronic access key (dongle). This paper presents an overview of the project, demonstrations of sample test items from the test battery,

presentation of the test delivery and administration system, review of test score results and psychometric analyses, and plans for future enhancements and extensions. The Singapore CATs use the following psychometric procedures: selection of initial item from a random proficiency value near the center of proficiency distribution of the selected item bank, Rasch model calibration and proficiency estimation, and a stopping rule based on a minimum standard error or administration of a specified maximum number of items. Results for the mathematics and reading CATs are presented showing scale score population distributions, stopping rule exit criteria, item exposure distributions, and ability estimate and standard error curves across the item administration sequence. The paper presents summary recommendations for enhancements and extensions with the CAT tests and additional CAT research and validity investigations.
The CAT results are based on examinee samples of approximately 12,000 for the reading tests and 9,000 for the numeracy tests.

For further information: rickard@casas.org

Criterion-Related Validity of an Innovative CAT-Based Personality Measure

Robert J. Schneider, *PDRI*
Richard A. McLellan and Tracy M. Kantrowitz, *PreVisor, Inc.*
Janis S. Houston and Walter C. Borman, *PDRI*

This paper blends rigorous and innovative psychometric theory with a practical selection application. We used CAT principles to estimate examinees' personality trait levels through an iterative, IRT-driven, paired-comparison assessment process. The concept has its roots in Thurstone's (1927) Law of Comparative Judgment. Thurstone conceived of using a paired-comparison procedure to scale stimuli on an interval scale. The idea was that if interval scale personality assessment could be generated with a paired-comparison procedure, then measurement might be made more precise than that yielded by typical Likert-type personality scales, which arguably provide only ordinal level data. Stark and Drasgow (1998) developed an algorithm to implement this process based on Zinnes and Griggs' (1974) probabilistic unfolding model which, in turn, is based on (and extends) the work of Coombs (1950) and Thurstone (1927). Examinees select which of the two statements representing different levels of a personality trait are more descriptive of them, and are then presented with two additional statements, based on their previous selection. Sequences of statement-pairs are selected in a manner that maximizes information in an IRT sense. Statement-pairs are presented for a given personality traits until either (1) a sufficiently low conditional standard error of measurement is reached, or (2) ten statement-pairs have been presented. This methodology has been used successfully in the Navy (Borman, et al., 2001; Houston, Borman, Farmer, & Bearden, 2005). To our knowledge, however, our measure represents the first commercial application of CAT to the personality domain. Our test measures thirteen traits selected to represent the broad personality sphere and to be predictive across a wide range of occupations and industries. Our intent was to build in flexibility to create composites of scales relevant to a variety of different work populations to accommodate the differing needs of our clients.

This presentation reports initial validity results. Our CAT personality measure was administered to 1,607 first-line supervisors in eight organizations, each of whom was rated by his/her immediate supervisor. Sample sizes for predictor-criterion pairings ranged from n = 745 to 1,109. To identify a composite of scales relevant to the supervisory position, we conducted a relative weight analysis (Johnson, 2000) to identify the relative importance of each predictor based on its proportionate contribution to R^2. This procedure controls for multicollinearity among predictors by considering the unique effect of each predictor as well as its effect when combined with the other predictors. Six scales were identified and a weighted sum was computed. The estimated operational validity of the adaptive personality scale composite was .25 against an overall job performance criterion. Graphs showing validity coefficients associated with presentation of different numbers of statement-pairs will also be shown for each scale included in the personality composite, as well as for the composite itself. This information will be very useful in that it will indicate how many statement-pairs must be presented to reach stable (asymptotic) criterion-related validity estimates.

For further information: Robert.Schneider@pdri.com

Computerized Adaptive Testing in Spain: Description, Item Parameter Updating, and Future Trends of eCAT

Francisco J. Abad and David Aguado, *Universidad Autónoma de Madrid*
Juan Ramón Barrada, *Universidad Autónoma de Barcelona*
Julio Olea, Vicente Ponsoda, and Francisco J. Abad, *Universidad Autónoma de Madrid*

eCAT is a CAT developed and applied in Spain to assess English proficiency in Spanish speakers. The test was developed by psychometricians from the School of Psychology (Universidad Autónoma de Madrid) and the IIC (Engineering Institute of Knowledge). Psychometricians constructed the item bank and designed the adaptive algorithm. The IIC takes care of the marketing and control of the test delivery via the Internet. At this time, thousands of tests have been administered in the context of the personnel selection processes and for the assessment of undergraduate's language competences in several Spanish universities. In this presentation we will summarize the work done for the design and updating of the system. We will address four different aspects of eCAT: (1) test construction, including item bank design and calibration, adaptive algorithm, psychometric properties of the θ scores (reliability and validity), computerized reports, and software for web-based application; (2) main results of the application (descriptive study of θ scores, estimation errors, execution time and exposure rates); (3) analysis of parameter drift and its impact on the θ scores, assessed by means of a comparison between the estimates of parameters in the initial calibration sample and those obtained under eCAT ordinary operation; and (4) work in progress: item parameter updating, increasing the bank size using on-line calibration procedures, and calibrating a new bank of items to assess the level of English listening (eCAT-listening).

For further information: fjose.abad@uam.es

Twenty-Two Years of Applying CAT for Admission to Higher Education in Israel Naomi Gafni and Yoav Cohen, *National Institute for Testing and Evaluation,* Israel

This paper describes the use of CAT in higher education admissions in Israel. This includes (1) the English as a foreign language (EFL) CAT that has been used by various institutions of higher education for placement purposes for 22 years; and (2) the CAT version of the Psychometric Entrance Test (MIFAM), which has been in use for nine years as a higher education admissions tool for examinees with disabilities. Both applications run in parallel with paper-and-pencil test (PPT) versions. This presentation will focus on the specific procedures used to produce equitable scores across the two media as well as on examining the suitability of the CAT for examinees with disabilities. The paper discusses a host of practical issues that were encountered during conversion of the Psychometric Entrance Test (PET) to a computerized adaptive format. Issues that pertain to the meeting of content specifications, item exposure, item banks, item bank dimensionality, and equating, are identified and discussed in the context of evolutionary changes in the MIFAM program.

For further information: naomi@nite.org.il

Item Selection and Hypothesis Testing for the Adaptive Measurement of Change

Matthew Finkelman, *Tufts University School of Dental Medicine*
David J. Weiss, *University of Minnesota*
Gyenam Kim-Kang, *Korea Nazarene University*

In a paper presented at the 2007 GMAC CAT Conference, Kim-Kang and Weiss (2007, 2008) described a procedure for the adaptive measurement of change (AMC) for an individual examinee. In this procedure, a CAT is administered at Time 1 to an examinee and the final θ estimate from that CAT is used to begin a second CAT at Time 2 (a later point in time). The Time 2 CAT continues until the Time 2 95% confidence interval around its θ estimate does not overlap the Time 1 95% confidence interval; when this occurs "significant change" is said to have occurred for that examinee. Kim-Kang and Weiss compared the performance of the AMC procedure in measuring change with that of change scores from conventional tests based on raw difference scores, residual change scores, and IRT-based difference scores. Their results showed that AMC captured change better than all methods based on conventional tests under a variety of test configurations and levels of true change. They also demonstrated that the AMC procedure was efficient in detecting significant change, requiring an average of from 6 to 22 items for different levels of true change.

The present study focused on the detection of change. Two new methods for testing the hypothesis of significant change for a single person were developed and compared to the confidence interval overlap approach. These methods were a likelihood ratio test approach and a Z-test approach. The power and alpha level of these two hypothesis testing methods were evaluated in the context of two CAT item selection methods—Fisher information and a variation of Kullback-Leibler information designed to select items in the context of AMC. The new methods were evaluated under subsets of conditions

examined by Kim-Kang and Weiss. Results demonstrated that both the likelihood ratio and the Z-test method had better control of alpha error and had better power to detect smaller amounts of change than the confidence interval overlap method. Item selection method had minimal effect on either alpha or power, with a slight difference in favor of AMC-modified Kullback-Leibler information. The combination of Kullback-Leibler information and the Z-test provided slightly better results than other combinations. When used with variable-length CATs, the latter combination resulted in substantial reductions in test length at Time 2 while maintaining alpha levels and power comparable to Time 2 fixed-length CATs. Recommendations are made for the further development of the AMC procedure.

For further information: mattstat2000@yahoo.com

A Gradual Maximum Information Ratio Approach to Item Selection in Computerized Adaptive Testing

Kyung (Chris) T. Han, *Graduate Management Admission Council*

One of the most widely used item selection methods in CAT is that of selecting an item with the maximized Fisher information (MFI) at the interim θ estimate based on previously administered items to the examinee (i.e., finding item x maximizing $I_x[\hat{\theta}_{m-1}]$ for an examinee with the interim θ estimate $\hat{\theta}$ and m-1 as the number of items administered (Weiss, 1982). However, interim θ estimates in the beginning of testing (e.g., before at least five items are administered) are rarely accurate, so applying the MFI method in the beginning of testing might not be the most efficient method and it might cause excessive exposure of those items with greater information. This study proposes a new approach in which the ratio between potential maximum information and expected information with an interim $\hat{\theta}$ is used as an item selection criterion in the earlier stages of CAT administration. The new approach, hereafter referred to as the gradual maximum information ratio (GMIR) approach, can be expressed as

$$\frac{I_x[\hat{\theta}_{m-1}]}{\max[I_x]}\left(1-\frac{m}{M}\right)+I_x[\hat{\theta}_{m-1}]\frac{m}{M}=I_x[\hat{\theta}_{m-1}]\frac{M-m(1-\max[I_x])}{\max[I_x]M}$$

where $\max[I_x]$ is the maximum information of item x (when θ is equal to the item difficulty), M is the test length, and m is 1 plus the number of items administered so far. Thus, using the GMIR method, those items that exhibit larger $I_x[\hat{\theta}_{m-1}]/\max[I_x]$ at the beginning of testing (i.e., when m is small) are more likely to be selected. Therefore, the GMIR approach is expected to utilize the whole item pool more efficiently and effectively in the earlier stages of testing and spare those items with greater information for the later part of testing.

A series of simulation studies were conducted to evaluate the effectiveness of the GMIR approach. The simulation studies mimicked one month of the existing CAT program for higher education (with

simplified content balancing and item exposure control) and used the evaluation criteria of item exposure rate, test information, item pool usage, and θ estimation bias and error (with root mean squared error). Another series of simulation studies were also conducted with the MFI method and the alpha stratification method (van der Linden, 2005). Each simulation study was replicated 100 times. The preliminary results showed that the GMIR approach very effectively utilized the item pool, providing satisfactory test information.

For further information: khan@gmac.com

Item Selection With Biased-Coin Up-and-Down Designs

Yanyan Sheng, *Southern Illinois University at Carbondale*

A basic ingredient in computerized adaptive testing (CAT) is the item selection procedure that sequentially selects and administers items based on a person's responses to the previously administered items. For decades, maximum information (MI; Lord, 1977; Thissen & Mislevy, 2000) has been widely used as the conventional algorithm for item selection in CAT. However, this criterion based on Fisher's information only targets the middle difficulty level where a person has about 0.5 probability of getting the items correctly, and hence is not applicable in situations where a different percentile is desired. In addition, MI heavily relies on an accurate estimation procedure that works well in all testing situations. Nonetheless, studies have shown that such a procedure is not readily available.

The biased-coin up-and-down design (BCD; Durham & Flournoy, 1994) has been widely used in bioassay for sequential dosage level selection because it can target any arbitrary percentile in addition to being efficient (Bortet & Giovagnoli, 2005). As the problem in bioassay shares many similarities with CAT, it is reasonable to believe that the item selection algorithm based on the BCD, which does not rely on an accurate trait estimate in every step of CAT administrations, provides an efficient alternative to, while being more flexible than, the conventional method. The development of this selection algorithm is essential as schools, professional organizations, and private companies seek to make CAT flexible enough to be implemented in wider testing applications.

The purpose of this study was to illustrate the use of the BCD in CAT and further evaluate its utility by comparing it with the conventional MI algorithm. For ease of comparisons, this study focused on the 1-parameter item response function. To investigate the utility of the BCD in CAT, two Monte Carlo simulation studies were conducted where either a fixed- or a random- stopping rule was employed. With fixed-stopping rule, the number of items administered was manipulated (k = 5, 10, 30, 100) and the item pool was fixed to have 100 different difficulty levels, whereas with random-stopping rule, the number of different difficulty levels in the item pool was manipulated (n = 10, 30, 50, 100). In either case, CAT responses were simulated for persons whose actual trait levels were 0 (average), −1 (1 standard deviation below the average), and −2 (2 standard deviations below the average), and the target difficulty level was at the 20^{th}, 50^{th} or 80^{th} percentile. Each adaptive testing simulation began the trait estimation with an initial value of 0 and proceeded with the maximum likelihood method. The results suggested that item selection with the BCD is more flexible in targeting any arbitrary percentile of the difficulty levels. With respect to the accuracy of the trait estimation, MI performs slightly better with fixed-stopping rule, whereas the BCD is considerably better for tests with small number of

different difficulty levels or persons whose trait levels are not at the extremes with random-stopping rule.

For further information: ysheng@siu.edu

A Burdened CAT: Incorporating Response Burden with Maximum Fisher Information for Item Selection

Richard J. Swartz, *The University of Texas M. D. Anderson Cancer Center*
Seung W. Choi, *Northwestern University Feinberg School of Medicine*

Widely used in various educational and vocational assessment applications, CAT has recently begun to infiltrate the patient-reported outcomes (PRO) arena. Several differences exist between PRO-CAT and "achievement CAT." Polytomous, rather than binary, items are more appropriate for PROs; constructs are often quasi-traits with skewed distributions; informative items cannot always be generated along the important range of the trait; and in many patient populations conditions exist so that patients cannot tolerate longer tests. Reducing this response burden has been one of the main reasons for consideration of CAT in the PRO arena. Although successful in reducing burden, many of the current CAT algorithms do not formally consider patient or examinee burden as part of the item selection process. In the PRO setting, many CAT applications simply limit the maximum number of items to be administered. This study uses a loss function approach motivated by decision theory to develop an item selection method that incorporates burden into the Maximum Fisher's Information (MFI) item selection method.

We compared several different loss functions representing varying degrees of burden, including a no-burden condition as a baseline. An item bank of 62 polytomous items measuring depressive symptoms was used to compare the different methods. The items were calibrated with the graded response model using 730 patients and caregivers from the M. D. Anderson Cancer Center. For each condition, we used two different response datasets to simulate CAT instruments. One dataset consisted of the real responses from the 730 patients and caregivers who answered all the items. The second dataset consisted of simulated responses to all the items based on a grid of θ values with replicates at each grid point. The MFI-burden algorithm for item selection results in tests that are on average shorter (depending on the degree of burden) than those obtained using MFI alone, but without severely affecting the standard error of measurement. In particular the loss function incorporating burden protects respondents from receiving longer tests when their estimated trait score falls in a location where there are few informative items. This is very useful in PRO assessment where burden to the patient is a concern.

For further information: rswartz@mdanderson.org

Adaptive Item Calibration: A Simple Process for Estimating Item Parameters Within a Computerized Adaptive Test

G. Gage Kingsbury, *Northwest Evaluation Association*

The characteristics of CAT change the characteristics of the field testing that is necessary to add items to an existing measurement scale. The process used to add field test items to a CAT might lead to scale drift (van der Linden & Glass, 2000; Ban, et al, 2001). In addition to this measurement concern, adding randomly chosen field test items to a test might disrupt the performance of an examinee by administering items of inappropriate difficulty. The current study makes use of the transitivity of examinee and item in IRT to describe a process for adaptive item calibration. In this process an item is successively administered to examinees whose ability levels match the performance of a given field test item. By treating the item as if it were taking an adaptive test, examinees can be selected who provide the most information about the item at its momentary difficulty level. Throughout the calibration process, the momentary difficulty estimate is updated and used in the process of item selection for all examinees. The item calibration can be completed when a fixed number of examinees have seen the item of interest, or when the momentary difficulty level for the item stabilizes to a predetermined variability. This approach should provide a more efficient procedure for estimating item parameters. While the procedure is not specifically designed to create an optimal calibration sample in the manner described by Holman and Berger (2001), it should result in the item being administered to a set of individuals that more closely approximates optimality.

The process is described in detail within the context of the one-parameter logistic IRT model. The process is then simulated using 10 replications of the calibration of 100 items to identify whether it produces more accurate and efficient item parameter estimates than random presentation of field test items to examinees. Results indicate that adaptive item calibration is more accurate for small sample sizes. With additional research, adaptive item calibration might provide a viable approach to expanding item pools in settings with small sample sizes or settings with a need for large numbers of items.

For further information: gage.kingsbury@nwea.org

On-the-Fly Item Calibration in Low-Stake CAT Procedures

Sharon Klinkenberg, Marthe Straatemeier, and Han van der Maas, *University of Amsterdam*

We present a new model for computerized adaptive progress-monitoring. This model is used in the Math Garden, a web-based monitoring system, which includes a challenging web environment for children to practice arithmetic skills. The Math Garden is a CAT web application, which tracks both accuracy and response time. Using a new model (Maris, in preperation) based on the Elo (1978) rating system and an explicit scoring rule, estimates of ability level and item difficulty are updated every trial. Items are sampled with a mean success probability of .75, making the tasks challenging yet not too difficult. By integrating the response time in the scoring rule, we try to compensate for the loss of information associated with the high success rates (van der Maas and Wagenmakers, 2005). In a period of eight months, our sample of 1,053 children completed over 850,000 arithmetic problems. The children completed about 25% of these problems outside their school hours. Results show good validity and reliability, high pupil satisfaction measured in playing frequency, and good diagnostic properties. The ability scores correlatde highly with the Dutch norm-referenced general math ability scale of the pupil monitoring systems of CITO. Also, test retest reliability analysis showed high

correlations. In view of the satisfactory validity and reliability of the person ability estimators, our method opens the door to on-the-fly item calibration in low-stakes testing.

For further information: S.Klinkenberg@uva.nl

An Automatic Online Calibration Design in Adaptive Testing

Guido Makransky and Cees. A. W. Glas, *University of Twente,* The Netherlands

An accurately calibrated item bank is essential for a valid CAT. However, in some settings, such as occupational testing, there is limited access to examinees for calibration. As a result of the limited access to possible xaminees, collecting data to accurately calibrate an item bank in an occupational setting is usually difficult. In such a setting the item bank can be calibrated online in an operational setting. This study explores three possible automatic online calibration strategies with the intent of calibrating items accurately while estimating ability precisely and fairly. That is, the item bank is calibrated in a situation where examinees' ability is assessed throughout the calibration design. The three calibration strategies represent a sample of possible designs on a continuum ranging from one extreme where items are calibrated at a single point in time, to the other extreme where items are calibrated constantly after each exposure. A simulation study was used to identify the optimal calibration strategy. The outcome measure was the mean absolute error of the ability estimates of the examinees participating in the calibration phase. Manipulated variables were: the calibration strategy, the size of the calibration sample, the item response mode, and the size of the item bank. The results of the study give an overview of the benefits of each strategy for different applied conditions, and provide viable calibration design options for test development companies that find it difficult to get examinees in the development phases of a test.

For further information: guidomakransky@gmail.com

Investigating Cheating Effects on the Conditional Sympson and Hetter Online Procedure with Freeze Control for Testlet-based Items

Ya-Hui Su, *University of California, Berkeley*

In CAT, if a group of examinees purposefully memorize items and distribute them to other prospective examinees, it certainly ruins the equality and accuracy of CAT. Steffen and Mills (1999) investigated this effect and found that the more the compromised items and the more effective the cheating, the more severe the overestimation for the recipients, especially for those with low ability levels. Su, Chen, and Wang (2004), pointed out that the overestimation for the recipients was more severe when the sources had diverse ability levels, because more items were compromised. Su and Wang (2007) proposed an item exposure control procedure, called the conditional Sympson and Hetter (Sympson & Hetter, 1985) online procedure with freeze control (denoted as SHCOF) procedure. Results showed it superior to many other conventional procedures in terms of measurement and operational efficiency. To assess the cheating effect, Su and Wang (2008) used the SHCOF procedure in a CAT, and found it could obtain precise estimation for persons in real time without requiring simulations to generate item exposure under a unidimensional context. In the past, little research has been done to investigate cheating effects within a testlet context. Hence, it is of great value to ascertain whether the SHCOF is also less affected

by the cheating between examinees under a testlet context, when compared to a popular procedure such as the conditional multinomial method (SLC; Stocking & Lewis, 1998). The goal of this study was to use simulations to investigate how these two item exposure control procedures would perform under various cheating conditions. It was hypothesized that SHCOF would be less affected by cheating than SLC.

Four independent variables were manipulated: (1) ability level of sources, (2) ability distribution of recipients, (3) cheating conditions (no cheating, inefficient cheating, efficient cheating, and perfect cheating), and (4) item exposure control procedure (SHCOF and SLC). The root mean squared error (RMSE) was computed to describe the cheating effects; the more serious the cheating effect, the larger the RMSE. Under the no-cheating condition, there is no significant difference in RMSE between SHCOF and SLC. It was also found that SLC had more serious inflation on RMSE than SHCOF under the perfect cheating condition. As the cheating condition got more severe, the overestimation for the recipients got more severe when the SLC was used. In addition, the more diverse the ability of the sources, the larger the RMSE and the mean positive bias would be. More importantly, SHCOF had smaller RMSE than SLC. This was because only SHCOF could simultaneously monitor item exposure and test overlap rates online. SHCOF could obtain precise estimation for persons without requiring simulations to generate item exposure before using in an operational CAT. If test items are memorized by sources and shared to recipients, CAT becomes unfair because the ability levels of the recipients will be overestimated. In this study, it was found that SHCOF was less affected by cheating than SLC. Hence, the SHCOF procedure can be safely implemented in operational CAT.

For further information: yahuisu@berkeley.edu

The Nine Lives of CAT-ASVAB: Innovations and Revelations

Mary Pommerich, Daniel O. Segall, and Kathleen E. Moreno, *Defense Manpower Data Center*

The Armed Services Vocational Aptitude Battery (ASVAB) is administered annually to more than one million military applicants and high school students. ASVAB scores are used to determine enlistment eligibility, assign applicants to military occupational specialties, and aid students in career exploration. The ASVAB is administered as both a paper-and-pencil (P&P) test and a CAT. CAT-ASVAB holds the distinction of being the first large-scale adaptive test battery to be administered in a high-stakes setting. Approximately two-thirds of military applicants currently take CAT-ASVAB; long-term plans are to replace P&P-ASVAB with CAT-ASVAB at all test sites. Given CAT-ASVAB's pedigree—approximately 20 years in development and 20 years in operational administration—,much can be learned from revisiting some of the major highlights of CAT-ASVAB history. This paper traces the progression of CAT-ASVAB through nine major phases of development including research and development of the CAT-ASVAB prototype, the initial development of psychometric procedures and item pools, initial and full-scale operational implementation, the introduction of new item pools, the introduction of Windows administration, the introduction of Internet administration, and research and development of the next generation CAT-ASVAB. A background and history is provided for each phase, including discussions of major research and operational issues, innovative approaches and practices, and lessons learned.

For further information: mary.pommerich@osd.pentagon.mil

The CAT-DI Project: Development of a Comprehensive CAT-Based Instrument for Measuring Depression

Robert D. Gibbons, *University of Illinois at Chicago*

The combination of IRT and CAT has proven invaluable in educational measurement. More recently, enormous reduction in patient and physician burden have been demonstrated using IRT based CAT in the area of mental health measurement problems (Gibbons et.al., 2008). CAT administration of a 626-item mood and anxiety spectrum disorder inventory revealed that an average of 24 items per examinee were required to provide impairment estimates with a correlation of 0.93 with the original complete scale. Furthermore, the CAT-based scores revealed twice the effect size than the total scale score in terms of differentiating patients with bipolar disorder based on the mood disorder subscale, despite an 83% reduction in the average number of items administered. These preliminary findings led to further interest and funding by the National Institute of Mental Health to develop a CAT-based instrument for the screening of major depressive disorder (CAT Depression Inventory—CAT-DI) that can be used for routine screening of depression in general medical practice settings as well as specialty mental health clinics. A recent supplement to the parent CAT-DI grant, extends our work on CAT for mental health measurement to CAT for diagnostic assessment of depression and other psychiatric disorders. The CAT Major Depressive Disorder (CAT-MDD) project will explore four different statistical/psychometric models for estimating the probability of an underlying discrete major depressive disorder based on self-administered symptom ratings that are adaptively administered. The ultimate objective of this program of research is to reduce patient and physician burden in terms of screening and diagnosing depression in general practice settings. Potential benefits include reduction in health care costs produced by high rates of service utilization among patients with an undiagnosed depressive illness, increased detection of depressive disorders, and increased access to quality mental health care for patients in need of such services.

For further information: rdgib@uic.edu

Development of a CAT to Measure Dimensions of Personality Disorder: The CAT-PD Project

Leonard J. Simms, *University of Buffalo*

In this presentation, describes the CAT-PD project, a funded, multi-year study designed to develop an integrative and comprehensive model and measure of personality disorder trait dimensions. Our general study aims are to (1) identify a comprehensive and integrative set of dimensions relevant to personality pathology, and (2) develop an efficient CAT method—the CAT-PD—to measure these dimensions. To accomplish our general goals, we plan a five-phase project to develop and validate the model and measure. The presentation describes the project generally, the results of Phase I (which is focused on content domains and initial item bank development), and our plans for IRT/CAT with these item banks. In particular, I will focus on how the item banks will be used, the possible IRT models we are considering for item bank calibration, the CAT algorithms we are planning to test, and our methods for deciding on a final set of procedures for the completed CAT-PD measure. Finally, I will discuss the CAT and IRT challenges that we anticipate facing in the future.

For further information: ljsimms@buffalo.edu

The MEDPRO Project:
An SBIR Project for a Comprehensive IRT and CAT Software System

The IRT Software

David Thissen, *The University of North Carolina at Chapel Hill*
and Scientific Software International

The IRTPRO (Item Response Theory for Patient-Reported Outcomes) component of the MEDPRO Project is an entirely new application for item calibration and test scoring using IRT. Fall, 2009 release of this software is anticipated; this presentation briefly describes its features, user interface, and output. IRTPRO provides maximum likelihood calibration of items fitted with the 1PL, 2PL, 3PL, Graded, Generalized Partial Credit, and Nominal IRT models in any combination, using one of three estimation algorithms: (1) Bock-Aitkin EM, (2) adaptive quadrature, or 3) Metropolis-Hastings Robbins-Monro (MHRM). Unidimensional or multidimensional IRT models might be used; among multidimensional models, the implementation performs full-information estimation for exploratory and confirmatory models, including the special-case treatment appropriate for bifactor models. Analysis of differential item functioning (DIF) is also provided, using the Wald test, with accurate item parameter error variance-covariance matrices computed using the Supplemented EM (SEM) algorithm. Several goodness-of-fit and diagnostic statistics are reported. Standard *maximum a posteriori* (MAP) and *expected a posteriori* (EAP) estimates of the latent variable(s) for item response patterns might be computed, as well as (weighted) summed-score to scale score translation tables.

For further information: dthissen@email.unc.edu

The CAT Software

Nathan A. Thompson**,** *Assessment Systems Corporation*

The CAT software for MEDPRO is designed to provide a comprehensive environment for the design and delivery of CATs. It consists of two main components: CATSIM and FASTCAT, in a package called CATPRO (Computerized Adaptive Testing for Patient-Reported Outcomes), which will be designed to interface with IRTPRO. CATSIM will be a major expansion of Assessment Systems' (ASC) POSTSIM software. CATSIM will implement post-hoc simulations, Monte Carlo simulations, and hybrid simulations of CATs. New features in CATSIM will include the addition of CAT for polytomous IRT models, item selection constraints (content balancing, item exposure controls and "enemy" items), and an expanded set of termination options. FastCAT will be an expansion of ASC's FastTEST Professional Testing System that includes all the options in CATSIM applied to the delivery of live CATs in a Windows environment. Output from both CATSIM and FastCAT will optionally be available in formats directly importable into IRTPRO for analysis and the parameter output from IRTPRO will be directly importable into both CATSIM and FastCAT.

For further information: nthompson@assess.com

Reviewing Test Overlap Rate and Item Exposure Rate as Indicators of Test Security in CATs

Juan Ramón Barrada and Julio Olea, *Universidad Autónoma de Barcelona,* Spain
Vicente Ponsoda, *Universidad Autónoma de Madrid*, Spain
Francisco J. Abad, *Universidad Autónoma de Madrid*, Spain

Test security is a major concern in CAT, because of the possibility of item sharing between examinees. A CAT will be considered more secure the lower the overestimation of the examinee's trait level due to item preknowledge. The common measures of test security have been the overlap rate between examinees and the distribution of item exposure rates. Usually, these indicators of test security have been evaluated when no item disclosure is present. We justify that lower overlap rates or less skewed distributions of usage of the items might not lead to safer CATs. The main ways of increasing security are to reduce: (1) the probability of item preknowledge of the first items administered, and (2) the overlap rate for high trait levels. In these conditions, there would be many different routes to obtain a high trait level estimation and it would be difficult for an examinee with item preknowledge to incorporate one of these routes. Progressive and proportional methods offer these characteristics. We show that these two methods are safer than the alpha-stratified method, a method with a much lower overlap rate. In fact, when the alpha-stratified method is applied, there is a "golden source of information:" an examinee with high trait level sharing items content is the best option for increasing trait estimation. When the progressive or proportional methods are applied, there is no source of information that fits to all the possible recipients. With these two methods, recipients and sources should have a similar trait level to lead to an important increment of trait estimation.

For further information: juanramon.barrada@uab.es

Optimizing Item Exposure Control and Test Termination Algorithm Pairings for Polytomous Computerized Adaptive Tests With Restricted Item Banks

Michael Chajewski and Charles Lewis, *Fordham University*

Much of the IRT and item exposure control literature regarding CAT has focused on the assessment of the impact of exposure control algorithms on frequency of item use, estimation precision, test bias, and overlap as well as item pool utilization and observed root mean square error rates. However, most inquiries into these pertinent issues have limited their inquiries to fairly large educational assessment-based item bank situations, which are less common in other areas into which CAT has been expanding. This paper discusses the results of a simulation study that focused on the pairing of item exposure control algorithms and test termination criteria within the specific framework of polytomous CATs using restricted item banks. Based on prior comparative and exploratory research by Chang and Twu (1998), Revuelta and Ponsoda (1998), Pastor, Dodd and Chang (2002), French and Thompson (2003), Davis (2002; 2004), Davis and Dodd (2005), Barada, Mazuelq and Olea (2006), Georgiadou, Triantafillou, and Economides (2007), and Barada, Olea and Abad (2008), six item exposure control algorithms and four test termination criteria were selected. Item exposure controls included the progressive-restricted maximum information method, Stocking and Lewis conditioning on estimated ability, target exposure control (TEC), Sympson-Hetter conditional strategy (SHC), 0-1

α-stratified strategy (0-1STR), and the combined α-stratified Sympson-Hetter method (STR-SH). The impact of these six algorithms was evaluated in their optimization of small item bank adaptive instruments using fixed length or fixed standard error (or Fisher target information) test termination criteria. Just like educational large test item bank assessments, restricted-item bank CATs also face issues regarding test security. Item exposure control algorithms are used to ensure limitations on any given item being delivered too many times. Non-cognitive assessments, which might also be high stakes, face an even greater need for test security since there are fewer items available. Alternatively, non-high-stakes instruments might need to utilize item exposure control algorithms for content validity purposes. Results are discussed in the framework of restricted item bank CATs such as non-cognitive psychological assessments and consumer survey evaluations.

For further information: chajewski@fordham.edu

Limiting Item Exposure for Key-Difficulty Ranges in a High-stakes CAT

Xin Li, Kirk A. Becker, and Jerry L. Gorham, *Pearson VUE*
Ada Woo, *NCSBN*

Item exposure control has become a critical and practical issue since CAT was widely implemented in test administration. Strategies for controlling item exposure have been developed to prevent overexposure of items while maintaining measurement precision. Randomization and conditional selection are two major types of exposure control techniques (Way, 1998). Randomization procedures allow a random component for controlling item exposure. Kingsbury and Zara (1989) proposed the "randomesque" method that randomly selects one item out of a prespecified number of the most informative items throughout the testing. Another method designed by Lunz and Stahl (1998) randomly selects from all items within a logit range of the optimal item difficulty. Alternatively, conditional selection strategies impose an exposure control parameter for each item given it is selected. The Sympson-Hetter method developed by Sympson & Hetter (1985) and modifications of this procedure are reviewed in Georgiadou, Triantafillou and Economides (2007). The most recently being presented by Barrada, Veldkamp and Olea (2009) is the multiple maximum exposure rate (r^{max}) method which defines as many values of r^{max} as the number of items. Chang and Ying (1999) also proposed an *a*-stratified CAT to limit the exposure of items with high discrimination by restricting their selection until θ estimates have stabilized. While adaptive tests using the Rasch model do not have exposure issues due to the item discrimination parameter, there can be problems with exposure for certain ranges of item difficulty. A Rasch-analog of *b*-stratified adaptive testing to control exposure in a key-difficulty range was investigated in this paper.

Numerous studies have been conducted to evaluate the effectiveness of a variety of algorithms that modify the CAT selection process to control item exposurel. Their strengths and weaknesses have been discussed for different models using dichotomous scoring, polytomous scoring, and testlet-based CATs. However, no studies have focused on exposure of items within a particular range, especially those items with difficulty level near the cut-score on variable-length adaptive tests. The CAT algorithm tends to overly administer these items under maximum item information selection. Overexposure of items might affect item parameter estimates and potentially the integrity of the test.

This research investigated multiple methods for limiting exposure of items near the cut score and evaluatde the results for measurement precision. Response data from a large-scale live CAT licensure exam were used to obtain the known item parameters for simulation. θ s for simulees were distributed according to the population distribution of final θ estimates on the live test. Four procedures were employed for controlling exposure of items near the cut score in a CAT, including the Kingsbury-Zara, the "within-.10-logits," the r^{max} method, and a stratified-*b* method. They were compared to a baseline condition with no exposure control. The performance of these procedures was evaluated first for measurement precision by the standard error of measurement. Other variables associated with test security include exposure rates, utilization of the item pool, and items overlap across test administrations.

For further information: Xin.Li@Pearson.com

Comparison of Adaptive Bayesian Estimation and Weighted Bayesian Estimation in Multidimensional Computerized Adaptive Testing

Po-Hsi Chen, *Taiwan Normal University*

The goal of the research was to compare two new Bayesian estimation methods, the adaptive Bayesian estimation and weighted Bayesian estimation, in multidimensional computerized adaptive testing (MCAT). Monte Carlo simulation and a multidimensional item response model, the multidimensional random coefficients multi-nominal logit model (Wang, Wilson, & Adams, 1997), were used in this research. Ten to sixty items of two-dimensional CAT were used with adaptive Bayesian, weighted Bayesian, and traditional Bayesian estimation. The dependent variables were conditional bias and the root mean square error (RMSE). Results indicated that these two new Bayesian approaches resulted in less regression bias than the traditional Bayesian estimation; however, weighted Bayesian estimation was more stable than the adaptive Bayesian estimation. The applications and suggestions for use of weighted Bayesian estimation are addressed

For further information: chenph@ntnu.edu.tw

Comparison of Ability Estimation and Item Selection Methods in Multidimensional Computerized Adaptive Testing

Qi Diao and Mark Reckase, *Michigan State University*

The impetus of this research is the lack of guidelines for designing multidimensional computerized adaptive tests (MCATs). There has been some research on unidimensional CAT on the properties of ability estimation and item selection methods (e.g. Weiss & McBride, 1984; van der Linden & Pashley, 2000). However, in the literature on MCAT, most studies use a single ability estimation and item selection method because they focus on other aspects of adaptive testing (e.g. Li Ip & Fuh, 2008). The only study on a comparison of different ability estimation and item selection methods for MCAT is Tam (1992). But that was before most currently used methods (e.g. Segall, 1996; Veldkamp & van der Linden, 2002) were developed. Also, most of the research has used two-dimensional cases, but we believe at least three dimensions are needed. In the proposed study, three ability estimation methods were compared. The first is the general maximum likelihood method (Segall 1996). A problem when

maximum likelihood is used is that estimates of location are not finite when the number of test items is small. One solution offered in Reckase (2009) is fixed-step-size maximum likelihood. This method updates the estimates of ability location with a fixed increment when infinite estimates are encountered. The third method is Bayesian estimation (Segall 1996).

In the proposed study, four item selection methods were compared. The first is maximizing the determinant of the Fisher information matrix (Segall 1996). The second is minimizing the trace of the inverse of Fisher information matrix (Mulder & van der Linden 2008). The third is maximizing the decrement in the volume of the Bayesian credibility ellipsoid (Segall 1996). The last is maximizing the Kullback-Leibler information (Veldkamp & van der Linden 2002). The ability estimation and item selection methods conditioning were compared using different priors and test length. The item pool was simulated based on data from the Michigan Educational Assessment Program mathematics test for 7^{th} graders. Mean bias and mean squared error (MSE) were used as a measure of estimation precision. Test length of 20 and 50 were generated and results were compared. For testing the impact of priors on the Bayesian method, a multivariate normal distribution with mean **0** and an identity variance-covariance matrix as in the real MEAP 2005 data were used and final ability estimates were compared. The maximum likelihood estimation method did not perform well for the test length of 20. When test length was 50, the estimates were much better. The fixed-step-size maximum likelihood method fixed the problem of estimates not converging and the results were comparable to the Bayesian method. Bayesian estimates were regressed toward 0 because Bayesian estimates tend to be statistically biased toward the mean of the prior. The standard errors of the estimation were smaller than the maximum likelihood method. Maximizing the determinant of the Fisher information matrix and minimizing the trace of the inverse of Fisher information matrix were comparable. When Bayesian ability estimation was used, the performance of Kullback-Leibler information was slightly better than the Bayesian item selection method with the test length 20. Those two methods were comparable with test length of 50.

For further information: diaoqi@msu.edu

Multidimensional Adaptive Test: The Application of Kullback-Leibler Information

Chun Wang and Hua-Hua Chang, *University of Illinois at Urbana-Champaign*

In adaptive testing, items are selected sequentially to match the updated ability of the examinee. Numerous item selection algorithms for item pools calibrated under unidimensional IRT models have been well developed. However, the assumption of unidimensionality can be easily violated, especially when the test covers broad content areas. In the presence of multidimensionality, instead of obtaining *m* separate unidimensional ability estimates, multidimensional IRT (MIRT) that provides a *m*-dimensional vector estimate might be a better choice. Previous researchers have shown that this kind of simultaneous estimation of abilities from different dimensions yields more accurate estimates, since it takes into account the correlational structure of those abilities. Built on MIRT, multidimensional adaptive testing (MAT) can, in principle, provide a promising choice in ensuring efficient estimation of each ability dimension. Currently, two item selection procedures have been developed for MAT, one based on Fisher Information embedded within a Bayesian framework, and the other using

Kullback-Leibler Information. Since Fisher information extends to a matrix, instead of a single value in multidimensional ability space, item and test information are no longer independent of each other. Therefore, the nice additive property of FisheriInformation does not apply to MAT. Alternatively, Kullback-Leibler information remains a single value and thus keeps its additive property.

It is well-known that in unidimensional IRT, the second derivative of K-L information (also termed "global information") is Fisher information evaluated at θ_0. This paper first generalizes the relationship between these two types of information in two ways—the analytical result is given as well as the graphical representation to enhance interpretation and understanding. It is shown that the complete Fisher information matrix can be easily recovered from K-L information, and the diagonals of the matrix equate to the curvature of the K-L information curve, evaluated with respect to each dimension separately. Secondly, a K-L information index is constructed in MAT, which represents the integration of K-L information over all of the ability dimensions. In geometric interpretation, this index is analogous to the volume under the information surface when only two dimensions are considered. This paper further discusses how this index correlates with the item discrimination parameters. In the two-dimensional case, an analytical derivation shows that the size of the K-L information index depends largely upon the sum of the squared item discrimination parameters, which is also termed "multidimensional discrimination". The results would lay a foundation for future development of item selection methods in MAT which can help equalize the item exposure rate. Finally, a simulation study will be conducted to verify the above results. The connection between the item parameters, item K-L information, and item exposure rate is demonstrated for an empirical MAT delivered by an item pool calibrated under two-dimensional IRT.

For further information: cwang49@illinois.edu

Multidimensional Adaptive Personality Assessment: A Real-Data Confirmation

Alan D. Mead, Avi Fleischer, and Jessica D. Sergent, *Illinois Institute of Technology*

Although CAT was developed in the context of ability tests (Weiss, 1982), studies have since demonstrated the effectiveness of CAT for measuring attitudes and personality. For example, Koch, Dodd, and Fitzpatrick (1990) applied the rating scale model to a Likert-scale attitudinal questionnaire. The rating scale model (an extension of the one-parameter logistic model for polytomous data) was found to fit the data very well and, although they noted item pool issues, succeeded in measuring effectively. Other studies have found similar results for personality assessments, suggesting that perhaps half the items of an assessment are needed to achieve comparable reliabilities (Waller & Reise, 1989; Reise & Henson, 2000). However, one issue that has not been extensively treated in prior literature is the multidimensional nature of most personality assessments. Prior research has generally applied unidimensional CAT to individual scales. Segall (1996) presented a multidimensional CAT (MCAT) methodology where correlations between the factors could be leveraged to administer and score items even more efficiently. Mead, Segall, Williams and Levine (1997) described a Monte Carlo simulation of the adaptive administration of the 16PF Questionnaire (Cattell, Cattell, & Cattell, 1993; Conn & Rieke, 1994) using Segall's MCAT method. As in Segall's

simulation, the MCAT method was effective in allowing additional reductions in assessment length, beyond those typically encountered with unidimensional CAT. For example, overall assessment length could easily be cut in half with small decrements in scale reliabilities.

The purpose of the current study was to extend the results of the Monte Carlo simulation (Mead, et al, 1997) to real data. This study is important for two reasons. First, it is always important to show that simulated results generalize to actual use. Even more importantly, recent research on personality (research that specifically included the 16PF; Chernyshenko, Stark, Chan, Drasgow, & Williams, 2001) has suggested that traditional IRT models do not fit personality data well and might not be the most appropriate models (Stark, Chernyshenko, Drasgow, & Williams, 2006). If the IRT model is a poor fit to 16PF data, the Monte Carlo results will not hold for real data. On the other hand, if the real-data results replicate the simulation results, then we might assume that traditional IRT models fit 16PF data sufficiently well. We obtained archival data from the administration of the 16PF Questionnaire to approximately 5,000 individuals and the two-parameter logistic model was fit to the items using BILOG-MG 3.0. Segall's (1996) software was adapted to read the actual responses of the individuals for a real-data simulation. Results generally supported the use of MCAT with 16PF items. Correlations between actual 16PF scores and MCAT trait estimates were high (averaging .91 to .82) for MCAT tests shortened by up to 40–50% while shorter MCAT tests had moderate correlations (averaging .72 to .58). The presentation will also discuss results for the pool usage (about a third of the pool had exposures greater than 90%), efficiency for individuals with extreme scores, and practical considerations for adaptive personality assessment.

For further information: jsergent@iit.edu

A Comparison of Three Procedures For Computing Information Functions For Bayesian Scores From Computerized Adaptive Tests

Kyoko Ito, *Human Resources Research Organization*
Mary Pommerich, and Daniel O. Segall, *Defense Manpower Data Center*

CAT requires a pool of items that can yield reliable scores for a range of examinees without compromising security. One way to evaluate CAT item pools is to compare them in terms of their information functions. The score information for any test score y is defined as:

$$I\{\theta, y\} \equiv \frac{\left(\frac{d}{d\theta}\mu_{y|\theta}\right)^2}{Var(y \mid \theta)} = \left(\frac{\frac{d}{d\theta}\mu_{y|\theta}}{SE(y \mid \theta)}\right)^2,$$

where θ is ability (Birnbaum, 1968). The numerator denotes the slope of the regression of score y on θ, while the denominator is the standard error of y for a given θ.For scores from tests that usually vary from examinee to examinee, Lord (1980, p.156–157) suggested a formula to approximate the information function by conducting simulations and obtaining the numerator and denominator at three successive θ points (referred to as "local method" because it is based on three points). He noted, however, that the slope can be quite unstable. To reduce the instability, the number of successive θ levels can be expanded to five (Segall, Moreno, & Hetter, 1997; referred to as the

"quasi-local method"). Yet another method seems possible if one has sufficient evidence that a single linear function fits data over the θ range—i.e., one based on the least-square regression slope across the entire θ range, coupled with $SE(y|\vartheta)$ (referred to as the "global-slope method"). The authors' recent research compared the three methods for one type of score, i.e., the maximum likelihood estimator (MLE). The current study is a follow-up study to make the same comparisons for another type of score, the Bayes modal estimator (BME) with a normal prior. The BME with a normal prior should have higher information than the MLE, particularly at the tails, because of the addition of the squared inverse posterior standard deviation.

Item responses were generated using the three-parameter logistic (3PL) model and item parameters for a fixed number of simulees at each of 31 equally-spaced θ points between –3.0 and +3.0. The source of the item parameters was a 900-item CAT pool comprised of items that are currently used in operational administrations of a large-scale testing program. Throughout the simulation, these item parameters were treated as "true" item parameters that were known. The simulated CAT procedure matched the actual operational implementation of the CAT testing program, including Sympson-Hetter exposure control and maximum information item selection. Two factors—test length and sample size—were manipulated in the comparison of the three procedures: (1) Test length: 10, 15, 30, and 60 items; and (2) The number of simulees at each equally-spaced θ point (N_k): 100, 500, 1,000, and 2,000.The results from the MLE study indicate that generally the three methods yielded very similar information functions, although, not surprisingly, the degree of similarity tended to vary depending on test length and *N*. The BME study used Bayes provisional and final ability estimates, as opposed to MLE estimates throughout. Use of BME versus MLE during item selection and scoring could affect the sequence of items that are administered, which could, in turn, affect the amount of score information.

For further information: kito@humrro.org

Adaptive Computer-Based Tasks Under an Assessment Engineering Paradigm

Richard M. Luecht, *The University of North Carolina at Greensboro*

Assessment engineering (AE; Luecht, 2007, 2008a, 2008b; Luecht, Gierl, Tan, and Huff, 2006) is a highly structured way of designing constructs and building instruments and associated scales that measure those constructs. By using construct maps, evidence models, task models and templates, AE makes it possible to generate extremely large numbers of test forms with prescribed psychometric characteristics (e.g., targeted measurement precision). This paper presents an extension of AE to include *computerized-adaptive performance tasks* (CAPTs). In a traditional CAT, each item is selecting to maximize the measurement precision relative to a provisional estimate of some latent trait. CAT requires every item to be calibrated using an appropriate IRT model so that estimates of item difficulty (location) and other characteristics can be used in the item selection process. Under AE, task models and templates can generate large classes of items. In turn, individual items *inherit* the estimated psychometric characteristics of the task models and/or templates. A hierarchical Bayesian framework is used for calibration and to quantify uncertainty associated with the class of items sharing

estimated item parameters (cf. Glas and van der Linden, 2003) . With CAPTs, features or components of the task models and/or templates are *altered in real-time* to actually vary the task difficulty in a systematic way. By applying a maximum information criteria to an item generation algorithm scripted as part of an AE template, the task features can be selected to create *highly variable computer-based performance tasks (i.e., items) that effectively adapt themselves to the proficiency of the examinee*. In this sense, the ensuing performance task or items become semi-intelligent measurement agents. The theoretical foundations for CAPTs will be presented in the context of several measurement scenarios. This paper will also present the hierarchical Bayes calibration framework and algorithms for item generation.

For further information: Email: rmluecht@uncg.edu

Developing Item Variants: An Empirical Study

Anne Wendt, *National Council of State Boards of Nursing*
Shu-chuan Kao and Jerry Gorham, *Pearson VUE*
Ada Woo, *National Council of State Boards of Nursing*

Large-scale standardized tests have been widely used for educational and licensure testing. In CAT, one of the practical concerns for maintaining large-scale assessments is to ensure adequate numbers of high quality items that are required for item pool functioning. Developing items at specific difficulty levels and for certain areas of test plans is a well-known challenge. This study investigated strategies for varying items that can effectively generate items at targeted difficulty levels and specific test plan areas.

Earlier researchers (LaDuca, Staples, Templeton, & Holzman, 1986, Bejar, 1996) described item modeling as a construct-driven approach to test development that is potentially validity-enhancing. Earlier research focused on mirroring cognitive processes in answering surveys for psychological performance (Bejar, 1993; Embretson & Gorin, 2001; Embretson, 1999; Bejar & Yocom, 1991), with the intention of generating isomorphic items. For large-scale testing, some item generation models are more statistics-driven (e.g., Glas & van der Linden, 2003) and others are more content-driven (e.g., Bejar, Lawless, Morley, Wagner, Bennett, & Revuelta, 2003). Each item generation model provides templates that allow decomposition of knowledge or skills and identification of the key components that constitute meaningful new items.

This research was a pilot study for procedures that will be expanded systematically in the future. Each variant item generation model was developed by decomposing selected source items possessing ideal measurement properties and targeting the desirable content domains. As Table 1 shows, four models were proposed to generate item variants.

Table 1. Variant Item Generation Models

Model	Definition in Item Developing
Key	Change key
Stem	Change stem
Distractor	Change a distractor
Hybrid	Multiple changes

Using these models, 342 variant items were generated from 72 source items. Two sets of experimental data were collected from three pretest periods. Items were calibrated using the Rasch model. Initial results indicate that variant items show desirable measurement properties. Compared to an average of approximately 60% of the items passing pretest, 84% of the variant items passed the pretest criteria.

It is expected that the use of variant item generation models can make item development more cost-efficient and less labor-intensive. Most importantly, the characteristics of the new items seem to be better controlled and more predictable than the "standard" methods for developing items (item writing and item review). Though this research is based on specific licensure exams, the methodology of this study might be applicable to other testing programs.

For further information: awendt@ncsbn.org

Evaluation of a Hybrid Simulation Procedure for the Development of Computerized Adaptive Tests

Steven W. Nydick and David J. Weiss, *University of Minnesota*

The ideal CAT has a large item bank with a wide range of item difficulties; furthermore, in order for the test to provide equiprecise measurements, there must be items that provide sufficient information across the full range of θ (Weiss, 1982). Post-hoc simulations have been proposed as a means of fine-tuning a CAT for live administration; indeed, Gibbons, Weiss, et al. (2008) demonstrated that the results of post-hoc simulations well predict the outcomes of a live CAT. However, before examining CAT test characteristics (e.g., SEM) with a post-hoc CAT simulation, each examinee must have provided a response to each item in a bank. But if the item bank is very large (e.g., 1,000), it might not be reasonable to expect any examinee to respond to all the items without factors external to the trait (e.g., fatigue) affecting his/her score. Frequently, because they tend to be large, CAT item banks are calibrated using concurrent calibration methods, which estimate IRT parameters from an incomplete data matrix including a set of linking items (e.g., Kim & Cohen, 1998). This paper proposes and evaluates the performance of a hybrid simulation procedure for use in developing CATs that employs these sparse, concurrent-linking matrices. The hybrid procedure estimates θ for each examinee with the item parameters estimated from the sparse linking matrix in conjunction with the set of item responses for each examinee. Then, the θ estimate for each examinee is used with Monte Carlo simulation methods to impute the examinee's missing data, resulting in a complete response vector for each examinee—part real item responses and part imputed simulated data. A post-hoc simulation is then implemented with the hybrid response matrix.

Two IRT models were used—two- and three-parameter logistic. From a simulated data matrix of 620 items and 1,000 examinees, either two, four, five, or ten item/examinee blocks were selected, with 20 anchor items, and the remainder of the items and simulees divided randomly into groups. Then, responses were deleted to items not belonging to a simulee's group, resulting in data matrices with from 49% to 87% missing data. Parameters were estimated for both the matrix of full responses and the matrix of partial responses and θ was estimated for each simulee. The new estimates of θ and the estimated IRT parameters were then used to simulate new responses. POSTSIM (Assessment Systems Corporation, 2007) performed a fixed termination (40 items) and a variable termination (SEM $\leq$ .20) post-hoc CAT on each matrix. For both the fixed and variable termination criteria, the hybrid CAT with parameters estimated from the full matrix of responses (HFP) had accuracy close to that of the hybrid CAT with parameters estimated from the partial matrix of responses (HPP), yet it also had efficiency close to that of a CAT performed on the full matrix of responses (FFP). The HPP had correlations with the FPP full-test θ well into the .90s; HPP and FPP performed poorly only near the

limits of estimating the 3PL (80 items per group). These results suggest that meaningful hybrid simulations can be performed with sparse data matrices involving up to almost 80% missing/imputed data. The simulation results were replicated with a real data set.

For further information: nydic001@umn.edu

Computerized Adaptive Testing for Cognitive Diagnosis

Ying Cheng, *University of Notre Dame*

CAT is a new mode of testing that enables more efficient and accurate recovery of latent traits. Traditionally, CAT is built upon IRT models that assume unidimensionality. With the advances of latent class models (LCM) and an increasing number of applications of them in testing and measurement, an interesting question that arises is how to build a CAT based on a LCM. Tatsuoka (2002) and Tatsuoka and Ferguson (2003) established a general theorem on the asymptotically optimal sequential selection of experiments to classify finite, partially ordered sets. Xu, Chang and Douglas (2003) proposed two heuristics on the basis of Tatusoka's theoretical work in the context of CAT, one using Kullback-Leibler information (the KL algorithm) and the other using Shannon entropy (the SHE algorithm). This paper presents an application of the optimal sequential selection method, i.e., selecting items sequentially for examinees during CAT, which is built upon a class of partially-ordered LCMs (i.e., the cognitive diagnostic models). Two new algorithms are proposed: (1) posterior-weighted KL information or PWKL method, and (2) a hybrid algorithm (HKL) which considers not only the posterior but also the distance between latent classes. Two simulation studies, one using simulated item parameters, the other with parameter estimates from real data, show that the PWKL and HKL algorithms outperformed the KL and SHE algorithms uniformly. Finally, we built the link among the algorithms by establishing equivalence between the Kullback-Leibler-information-based approaches and the Shannon-entropy-based approach, and connecting the algorithms for LCM with algorithms built upon IRT models.

For further information: ycheng4@nd.edu

Obtaining Reliable Diagnostic Information through Constrained CAT

Jeff Douglas, Hua-Hua Chang, and Chun Wang, *University of Illinois at Champaign*

We consider how constraint weighted *a*-stratification can be used in CAT to guarantee that sufficient diagnostic information is obtained on a set of binary latent attributes, when estimation of a unidimensional IRT ability parameter is also desired. Such applications are useful when a single score is needed, but a more fine-grained assessment of the particular skills of an examinee is also desired. Accomplishing these dual aims requires carefully constructing how a single underlying model might simultaneously contain information about a continuous latent trait and a set of binary latent attributes of a cognitive diagnosis model. Such a model is discussed and results are given illustrating how these competing models can both be thought of as valid for an exam. Implementation of constraint weighted *a*-stratification involves identifying a priority function that combines IRT with cognitive diagnosis. Several priority functions are proposed, some based on formal measures of information, and others only utilizing knowledge of which items measure which attributes. A

simulation study and results are reported, showing how utilization of information-based methods yields higher classification rates for cognitive diagnosis while achieving accurate ability estimation. Item exposure rates are also considered for all competing methods. Several new directions for future research are proposed, both for item selection and for considering when multiple latent variable models for a single dataset can be simultaneously used to extract useful information.

For further information: jeffdoug@illinois.edu

Applying the DINA Model to GMAT Focus Data

Alan Huebner, Xiang Bo Wang, and Sung Lee, *ACT, Inc.*

Recent years have seen growing interest in the area cognitive diagnostic modeling. These relatively new psychometric models seek to classify examinees as having mastered or not mastered a set of discretely defined skills, as opposed to traditional IRT models that assign examinees a continuous score measuring a broadly defined latent trait. The literature in this field contains few examples of applications of cognitive diagnostic models to real assessment data, and many of these applications use simple datasets as a means of introducing a new estimation algorithm. We attempt to fit the Deterministic Input, Noisy-And (DINA) model to assessment data for an existing test, the GMAT Focus. We discuss whether useful diagnostic information can be gleaned by applying the model to the data.

For further information: Alan.Huebner@act.org

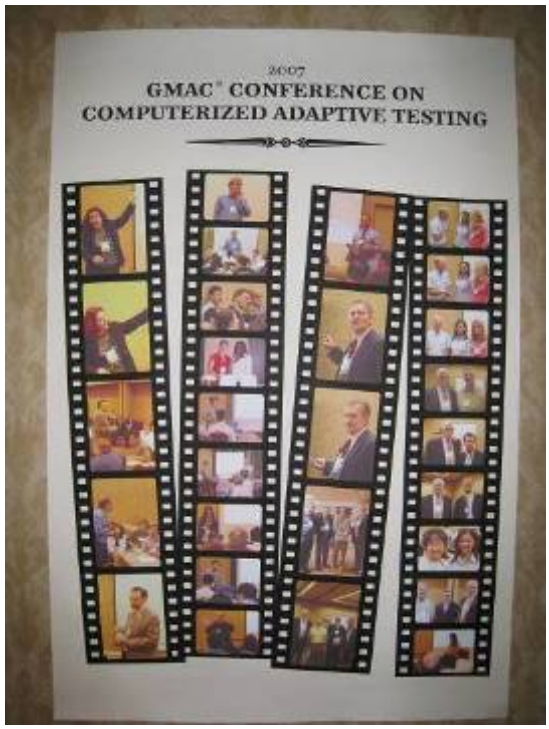
附錄三：照片集錦

H1N1 防護篇：起程

起程：搭乘輕軌車

住宿 Radisson University Hotel(三天)

會場標示圖



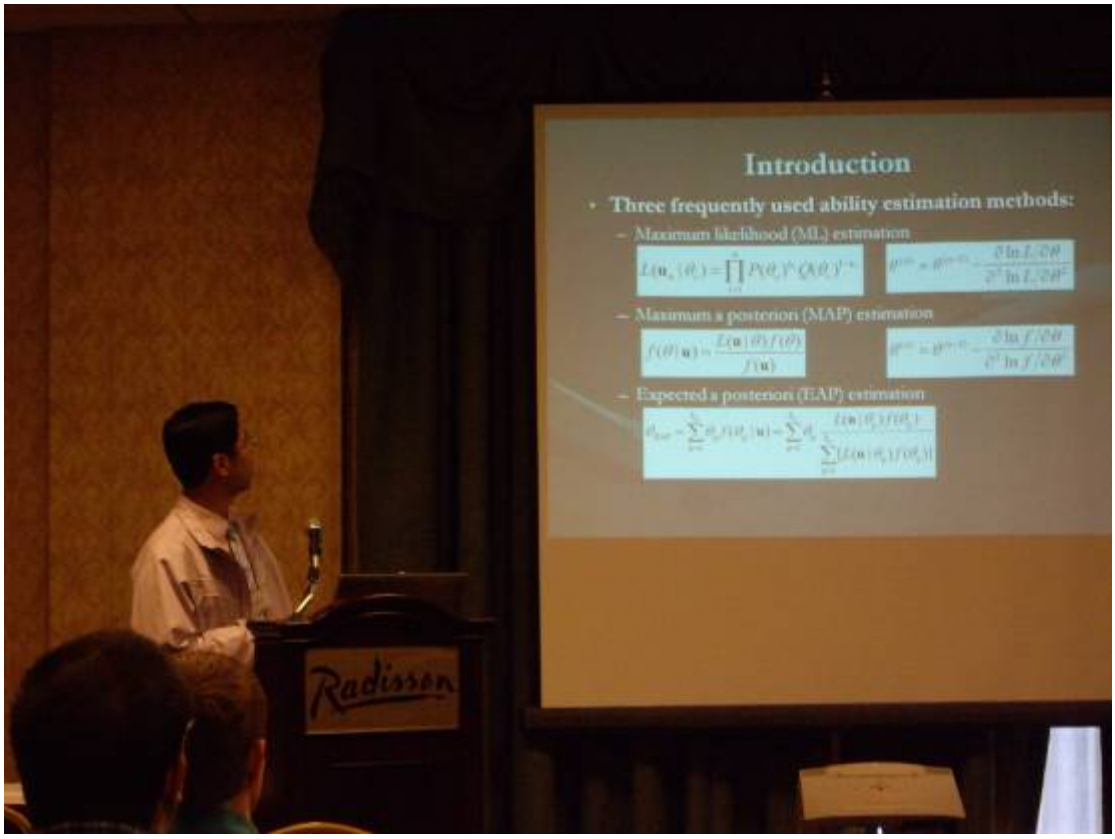
會場文宣：2007 年 GMAC CAT 與會者照片剪輯

會議行程：報到

會議行程：開幕

會議行程：會議議題

會議行程：會議議題



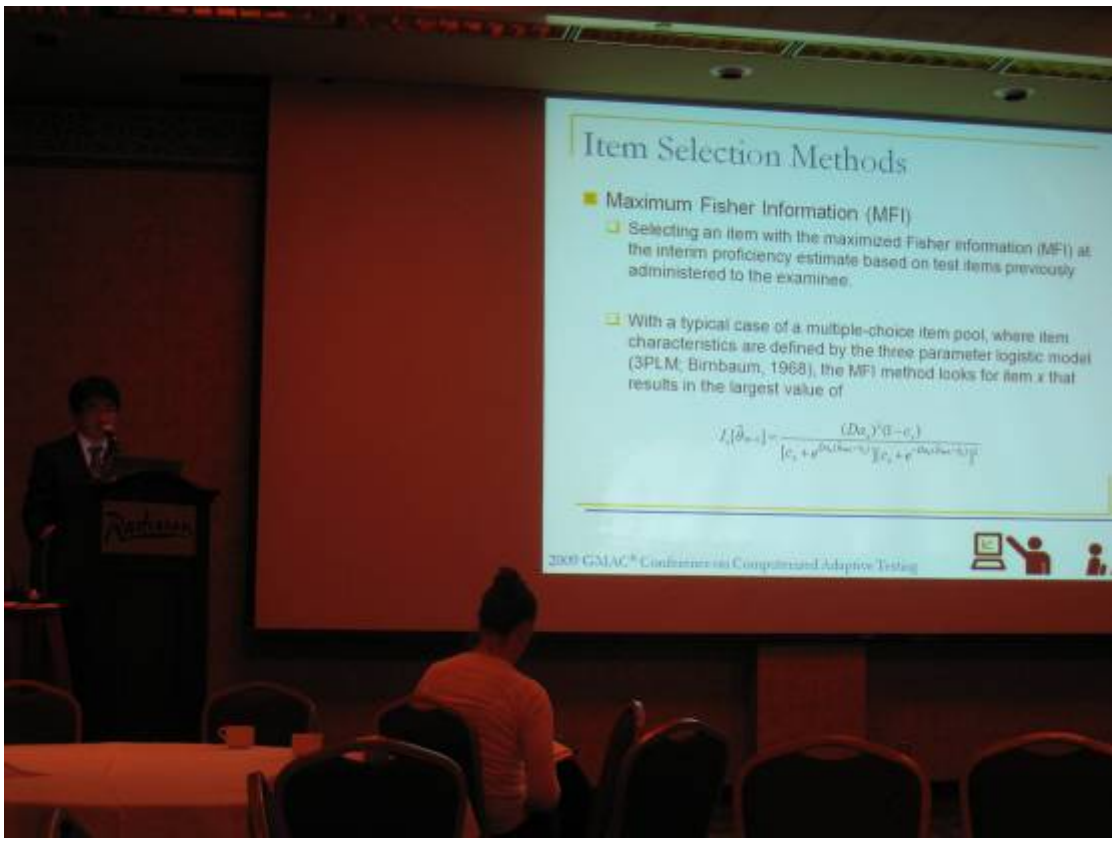
會議行程：會議議題(主講者：師大陳柏熹教授)

會議行程：會議議題

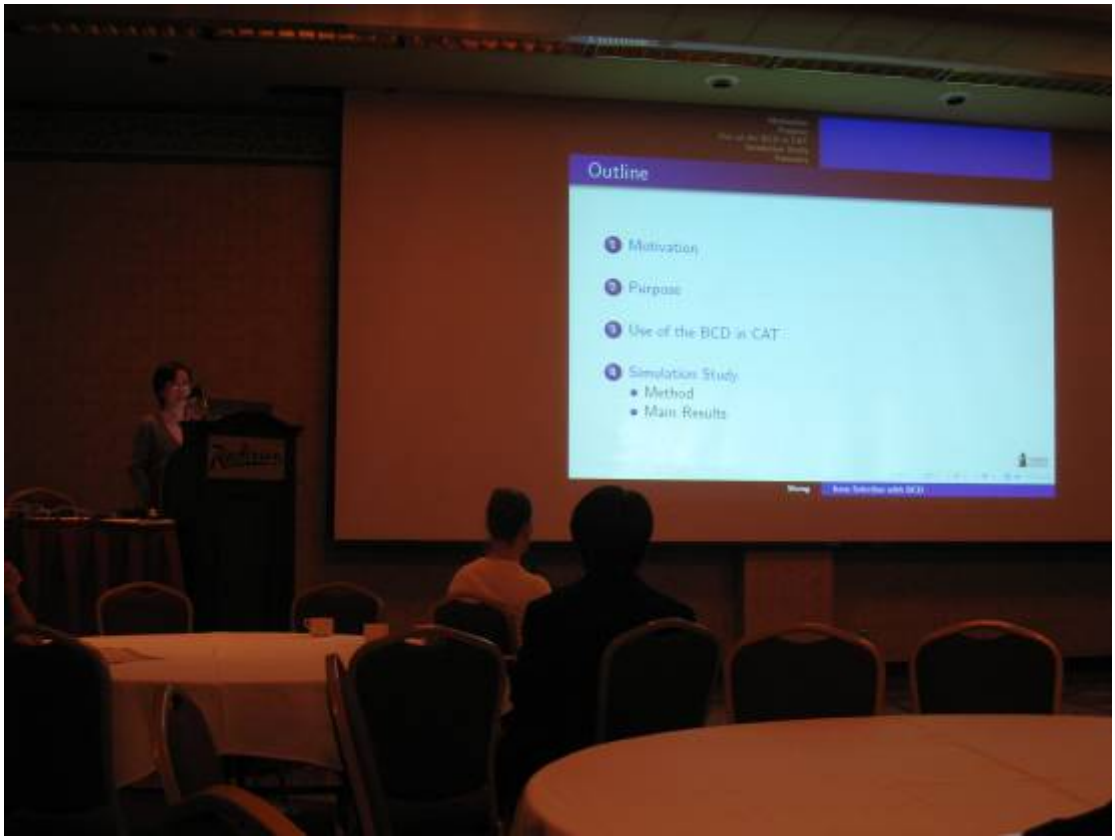
會議行程：會議議題

會議行程：會議議題

會議行程：會議議題



會議行程：會議議題



會議行程：會議議題

會議行程：會議議題

會議行程：會議議題

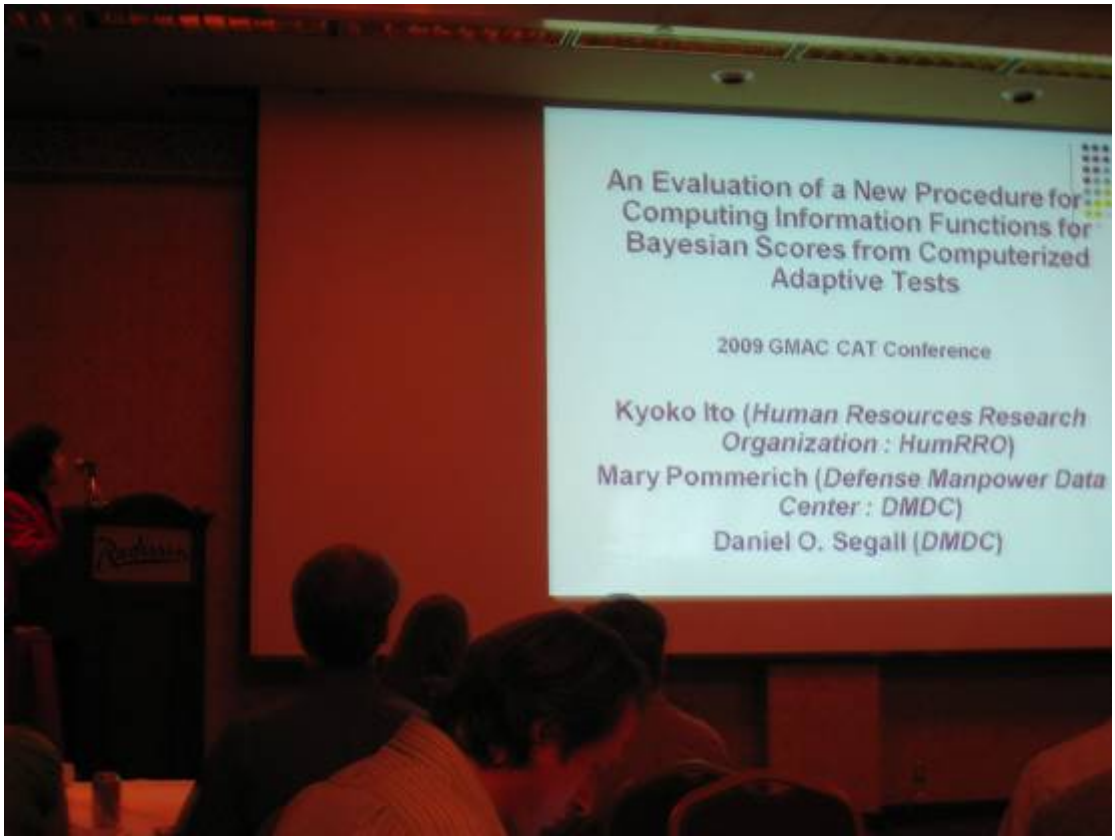
會議行程：會議議題

會議行程：會議議題

會議行程：會議議題

會議行程：會議議題

會議行程：會議議題



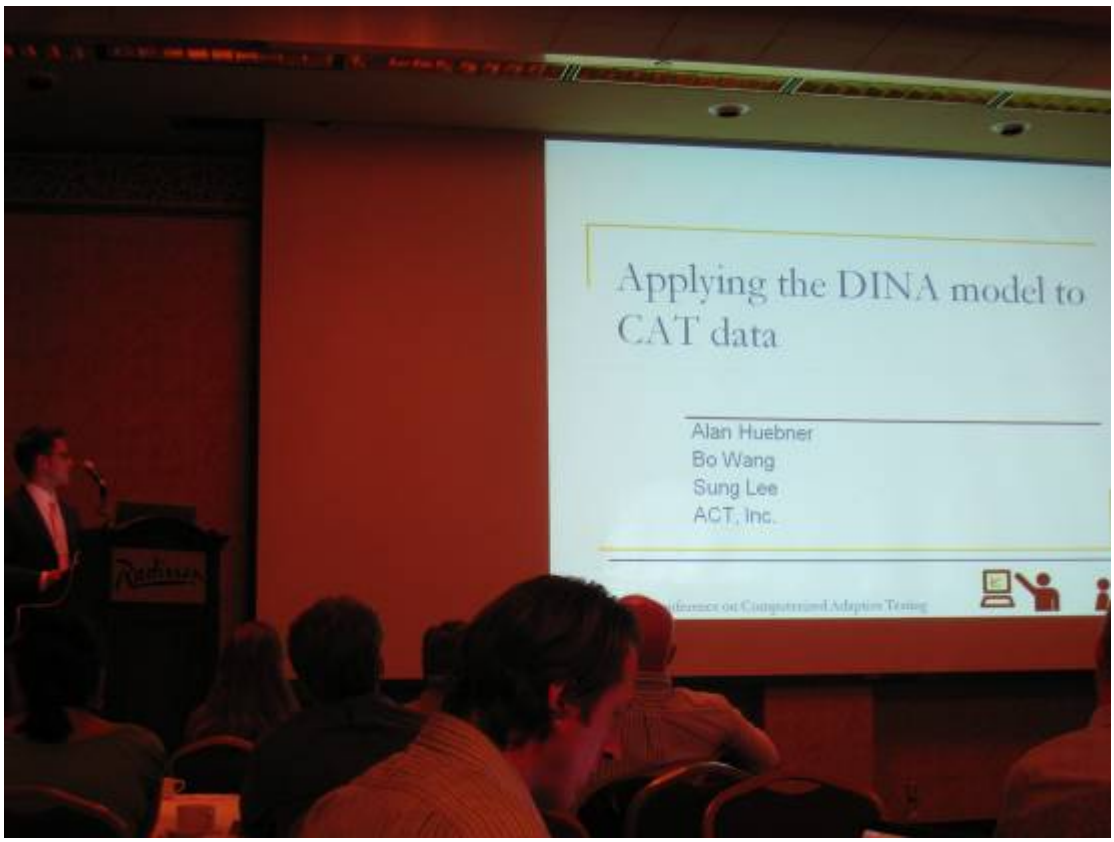
會議行程：會議議題

會議行程：會議議題

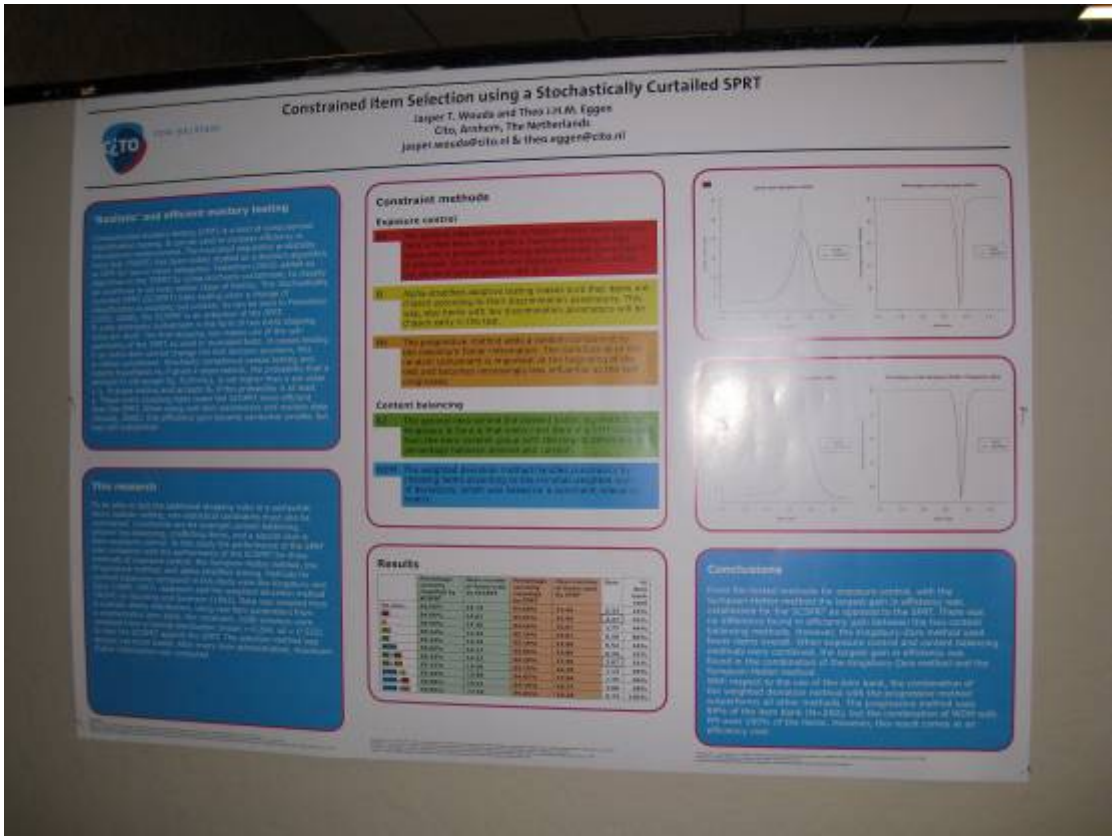
會議行程：會議議題

會議行程：會議議題

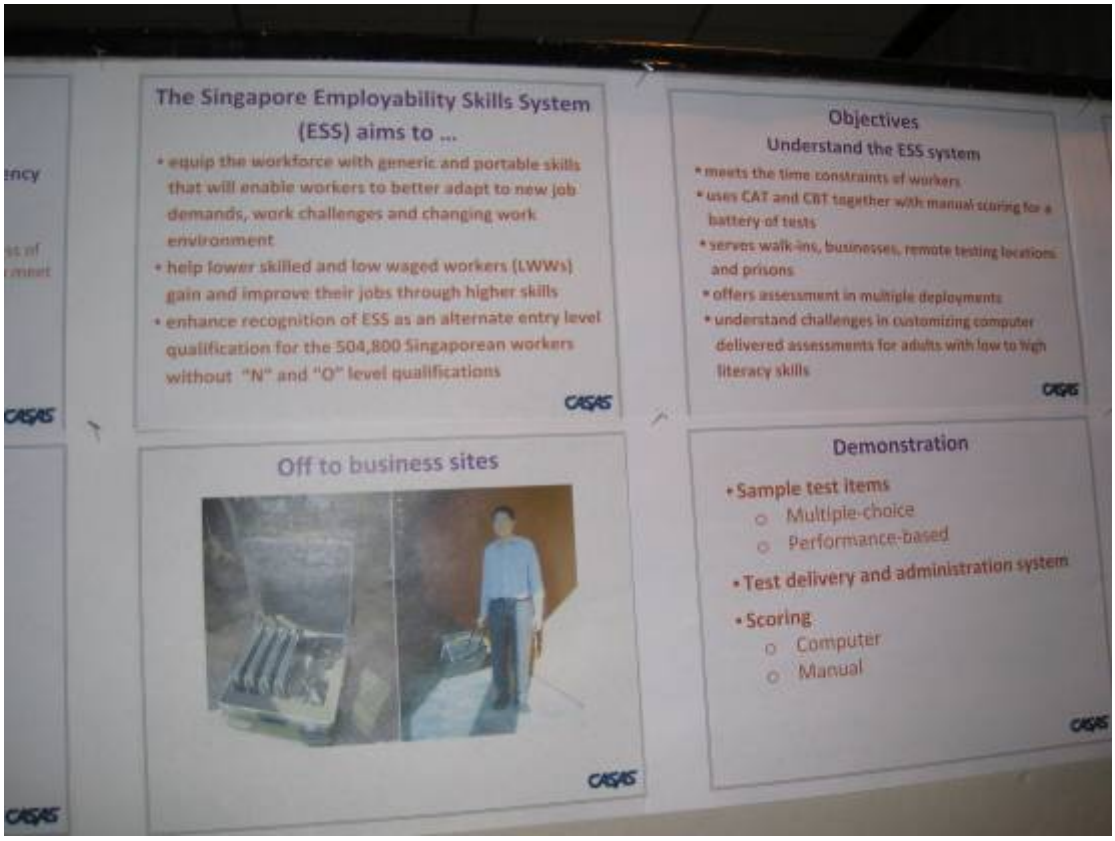
會議行程：會議議題



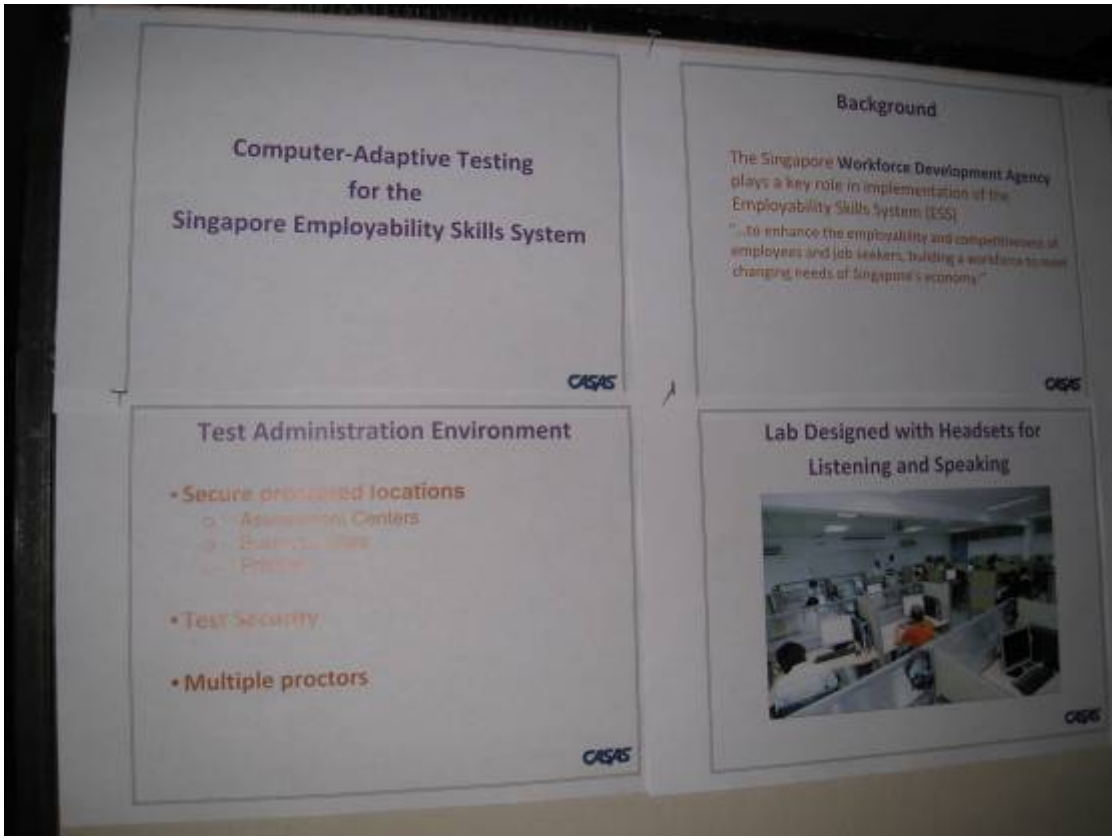
會議行程：會議議題



會議行程：POSTER SESSION



會議行程：POSTER SESSION



會議行程：POSTER SESSION

會議行程：POSTER SESSION

會議行程：POSTER SESSION 留影

友誼篇：與以色列主講者合影

飲食篇：午餐與陳柏熹教授合影

城市建設觀摩

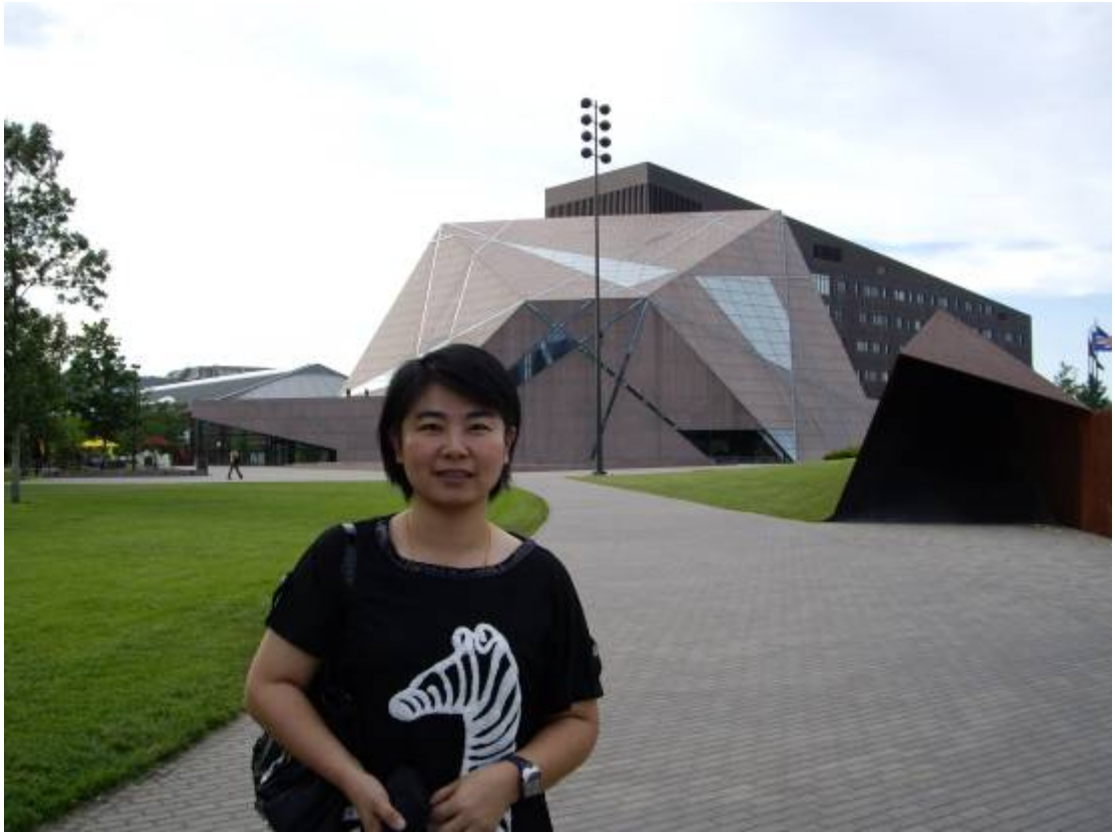
城市建設觀摩

城市建設觀摩

校園建築觀摩

校園建築觀摩

校園建築觀摩



校園建築觀摩


Permit Parking
Public Parking
Posted
MINNEAPOLIS PARK & RECREATION BOARD
East River Flats

校園觀摩

校園建築觀摩

H1N1 防護篇：返程