應用改良式株落選擇演算法同時組裝平行測驗

許祐福
彰化師範大學 數位學習研究所
碩士生
m96332008@mail.ncue.edu.tw

鄭培成
清雲科技大學 資訊管理系
助理教授
pccheng@cyu.edu.tw

江茂綸
朝陽科技大學 資訊與通訊系
助理教授
mlchiang@cyut.edu.tw

胡芸菁
彰化師範大學 數位學習研究所
碩士生
m98332008@mail.ncue.edu.tw

張庭毅
彰化師範大學 數位學習研究所
助理教授
tychang@cc.ncue.edu.tw

摘要

本文提出了一個改良型株落選擇演算法來同時組裝多份擁有相同測驗訊息量和測驗特性的 TCC/TSIF 平行測驗。該方法能將同時組裝多份測驗的問題簡化成單一測驗組裝問題，因此不會發生循序組裝的不平等問題，也不需要加入大量的選擇變數和限制式。經實驗結果指出，在題庫有足夠試題的情況下與基因演算法以循序方式組裝多份平行測驗相比，CLONAL 不但有較小的偏差，且能在各種條件下組裝符合 TCC/TSIF 的平行測驗。

關鍵詞：平行測驗組裝問題、同時測驗組裝、項目反應理論、株落選擇演算法、啟發式演算法

Abstract

This article proposes a novel method based on an improved CLONALG algorithm to simultaneously construct TCC/TSIF parallel tests which have the identical test information functions and test characteristic curves. This method greatly simplifies a simultaneous parallel test construction model into a single test construction model. At the same time, it avoids the inequality problem in the sequential construction and solves the drawback of a larger number of variables and constraints in the simultaneous construction. The experiment results show that the proposed method in simultaneously constructing parallel tests has a lower deviation than *Linear Programming* (LP) and the *Genetic Algorithm* (GA), and it can construct TCC/TSIF parallel tests efficiently in various test specifications.

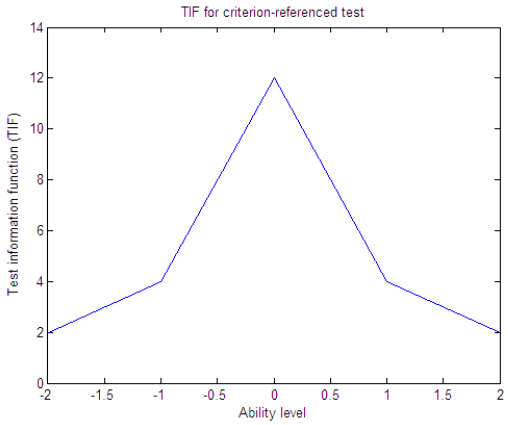
Keywords: parallel test construction problem, simultaneous test construction, item response theory, clonal selection principle, heuristic algorithm.

1. 前言

論測驗是教育領域中用來檢驗學生學習成效的重要工具，而試題反應理論(IRT)是近年來測驗領域中的主流技術，它已經被應用在各領域的大型測驗中，如托福、GRE、GMAT 考試，以及一些人格量表的編製。著名的電腦化適性測驗(CAT)也必須仰賴 IRT 的理論與技術才能運作。隨著應用層面越來越廣，組裝一份高品質的 IRT 測驗變得越來越重要。

而建構基於IRT的測驗主要包含三個步驟[1]，以本文所考慮的三參數對數模型為例[2]。第一，先將所有試題的鑑別度、難度、猜測度轉換成試題訊息量。第二，依照不同的測驗目的來設定目標測驗訊息量(TIF)，作為選擇試題的依據，而較常見的測驗目的包括效標參照測驗(criterion-referenced test)和廣泛能力測驗(board abilities test)。以效標參照測驗為例，其測驗的目的為區分出精熟者和非精熟者，因此TIF呈現單峰形曲線分佈，波峰位於指定的能力點上(見圖一)。最後，經由測驗組裝方法從題庫中選出試題的組合，選出的試題訊息量加總需等於目標TIF。由上述流程可知，一份高品質的IRT測驗除了仰賴設計完善的題庫(包含高鑑別度和各種難度的試題以及足夠的試題數)，還需要有高效能的測驗組裝方法，以便在合理的時間內，組裝滿足多評估準則的測驗規格。因此測驗組裝方法是測驗領域中的重要議

題，相關的研究也不曾中斷過[1, 3-14]。



圖一. 將受試者分成兩群的目標 TIF

而在實際測驗組裝時，經常需要一次組裝多份測驗。像是(1)週期性的測驗(2)在測驗時，給不同的受試者不同試卷以避免作弊(3)包含前測和後測的二階段甚至多階段測驗(4)多階段的適性測驗。這些擁有相似測驗規格的多份測驗就稱為平行測驗[15-16]。在 Theunisse 首次使用數學規劃模型來組裝 IRT 測驗時，是採用循序的方式來組裝平行測驗[17]。其作法是先組裝第一份測驗，在組第二份測驗前，先把第一份測驗的試題從題庫中移除，才進行第二份測驗的組裝程序，其後的第三份測驗所使用的題庫也剔除了第一份和第二份測驗的題目，以此方式確保每份平行測驗的試題不會重疊。循序組裝優點在於實作簡單，只要執行多次單一測驗組裝的程序即可，而測驗組裝時間隨著測驗的複本數呈線性增加。然而，循序組裝的每份測驗都在不平等的情況下挑選試題，因為優良的試題很容易被先組裝的測驗選走，越後面組裝的測驗可挑選的優良題目越少，導致測驗的品質下降，嚴重時甚至會發生無法組裝測驗的情況，這就是循序組裝產生的不平等問題[15, 18]。

為了解決不平等問題，Boekkooi 在測驗組裝問題的模型中加入試題不能重疊的限制條件，提出同時組裝多份測驗的概念[19]。以下便是同時組裝 F 份平行測驗的模型及其所使用的變數:

- x_{if}：二進制選擇變數，若為 1，表示第 f 份測驗選擇試題 i，反之則否
- N：題庫試題數
- F：平行測驗數
- $I_i(\theta_k)$：試題 i 在能力等級 k 的訊息量
- $T(\theta_k)$：能力等級 k 的目標訊息量
- n：目標測驗長度

$$\text{Minimize} \quad \sum_{f=1}^{F}\sum_{k=1}^{K}\sum_{i=1}^{N} I_i(\theta_k)x_{if} \tag{1}$$

S.t.

$$\sum_{i=1}^{N} I_i(\theta_k)x_{if} \geq T(\theta_k), \quad k=1,...,K, f=1,...,F \tag{2}$$

$$\sum_{i=1}^{N} x_{if} = n, f=1,...,F \tag{3}$$

$$\sum_{f=1}^{F} x_{if} \leq 1, i=1,...,N, f=1,...,F \tag{4}$$

$$x_{if} \in \{0,\ 1\}, i=1,...,N \tag{5}$$

限制式(2)和目標式(1)指出試題組合的訊息量須在大於等於目標訊息量的前提下，找出一個最小值，以此方式限制組裝測驗的TIF須滿足目標值。限制式(3)指出組裝測驗須符合目標長度。限制式(4)便是試題重疊的限制條件，指出一個選題變數x_{if}在F份測驗中只能被選擇一次。上述模型實現了同時組裝的概念，也避免了循序法在不同次組裝間所產生的不平等問題。然而在此模型中，僅僅是避免試題重疊的限制條件數量就等於N乘上F，隨著題庫試題數和平行測驗數的增加，大量的選擇變數和限制式會造成Linear Programing (LP)方法無法在可接受的時間內收斂。使得同時組裝的技術僅適用於較小型的測驗組裝問題[15, 20]。

面對同時組裝技術所遇到的瓶頸，Ackerman 回到循序組裝上作改良，提出兩階段的循序組裝方式[3]。其作法是，在第一階段以循序的方式組裝多份平行測驗，然後在第二階段將多份平行測驗的試題在不違反測驗規格的前提下進行互換，以進一步降低整體誤差。這樣的作法能稍微降低不平等現象所造成的誤差。延續 Ackerman 的兩階段組裝概念，Adema 將此概念應用在同時組裝的技術，提出一個二階段的同時組裝方法[4]。作法是在第一階段直接組裝一份大型測驗，該大型測驗是 F 份試卷的試題集合。因此其試題數為單份測驗的 F 倍，測驗訊息量也是 F 倍，若有其他的限制式也都乘上 F 倍，下列出此概念的問題模型:

$$\text{Minimize} \quad \sum_{k=1}^{K}\sum_{i=1}^{N} I_i(\theta_k)x_i \tag{6}$$

S.t.

$$\sum_{i=1}^{N} I_i(\theta_k)x_i - T(\theta_k)F \geq 0, k = 1,...,K \tag{7}$$

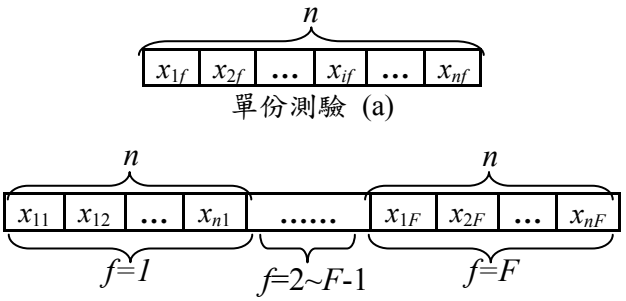
$$\sum_{i=1}^{N} x_i = nF \tag{8}$$

$$x_i \in \{0,\ 1\}, i = 1,...,N \tag{9}$$

在第二階段,使用啟發式演算法把在第一階段組裝的大型測驗拆分成F份規格相似的平行測驗。這樣的作法，等於是用單一測驗的模型去同時組裝多份平行測驗，因此既不會有循序法中不平等的問題，也不會有同時組裝中的變數和限制式過多的問題。然而，即便第一階段能組裝一份滿足測驗規格的大型試題集合，該試題集合所分割成的多份平行測驗也不一定能平均的拆分成F份規格相同的平行測驗，因此在平行測驗之間的誤差仍有很大的改善空間。

van 吸收了 Adema 以組裝一份大型測驗來代替多份測驗的概念，將此概念應用回循序法，提出虛擬測驗的概念來解決不平等問題[15]。虛擬測驗概念的特點在於，以循序的方式組裝測驗，但不會發生不平等的問題。因為該方法在每一次的循序組裝時，除了組裝一份目標測驗外，還會同時組裝一份剩餘測驗的總試題集合(虛擬測驗)。例如要組裝 5 份平行測驗，則在第 2 次循序組裝時，除了組裝第 2 份測驗外，還同時組裝一份包含第 3~5 份測驗的大型虛擬測驗。程序結束後只挑出第 2 份測驗，把包含第 3~5 份的虛擬測驗的試題全部放回題庫，以這樣的方式循序地組裝多份測驗。這樣的作法可確保較早組裝的測驗不會選走所有高品質試題。因為虛擬測驗也會取得相同比例的高品質試題，所以會有足夠的高品質試題流回題庫，供下一次的循序組裝程序使用。這樣就能降低不平等現象所造成的問題，而且採用循序的方式也比 Adema 方法的誤差來得低很多。然而，這樣的作法等於是以循序的方式進行了F次的同時組裝。只是解決了不平等問題，卻使得計算量大幅上升。

回顧過去的研究，只有少數研究能解決不平等問題且不需加入大量的變數和限制式，但這些方法不是測驗不夠符合目標規格，就是計算太繁瑣，沒有一種方法能在短時間內同時產生多份低誤差的平行測驗。本文的目的是要從本質上解決同時組裝所遭遇的變數和限制式過多問題，提出一個能同時組裝多份低誤差的平行測驗的方法。由於測驗組裝是一種組合最佳化問題，可能的試題組合將隨著試題數量增加而呈指數成長[1]。所以近年來的研究朝向運用各種啟發式演算法，以在合理的時間內組裝出可接受的測驗。像是類神經網路[7]、蒙地卡羅演算法[8]、禁忌演算法[10]、粒子群最佳化演算法[11]、免疫演算法[21]及基因演算法[9, 13-14]。雖然上述研究都是針對單一測驗的組裝，但值得注意的是，這些演算法只要經過改良就可以克服變數和限制式過多問題。像是黃採用實數型選題變數(試題序號)的禁忌演算法來同時組裝多份測驗[12]。其作法是在測驗編碼時，將多份測驗串在一起(見圖二)，其中x_{if}便是實數型的選題變數(試題序號)。



圖二. 測驗的表達(編碼)

這樣就不需要試題重疊的限制式，藉由修改演算法的特性就可簡化平行測驗的問題模型。然而，串接測驗會大大的增加測驗組裝問題的搜尋空間，導致組裝的測驗品質下降。舉例來說，假設有一個 4000 題的題庫，要組裝 3 份 40 題的平行測驗，因此N=4000，n=40，F=3。經由組合的公式可計算出兩種作法的試題組合總數量。循序組裝 3 份測驗的組合總數為 $F\text{x}C_N^n = 3\text{x}C_{4000}^{40} = 5.16\text{x}10^{36}$，而用串接的方式同時組裝 3 份測驗的組合總數為 $C_N^{F\text{x}n} = C_{4000}^{120} = 3.56\text{x}10^{244}$，搜尋空間足足增大了 $6.89\text{x}10^{207}$ 倍。因此該方法還是沒有解決同時組裝平行測驗所遭遇的高計算複雜度問題。

總結來說，同時測驗組裝的相關研究雖能有效的克服循序法的不平等問題[4, 12, 15, 19, 22]，卻大大提高了問題的複雜度導致平行測驗間的誤差提升。此外，以往的測驗組裝研究都著重於組裝 TSIF-Parallel tests(或稱弱平行測驗, weakly parallel tests)，而較少針對 TCC-parallel 的測驗組裝研究。根據 McDonald 的定義，TSIF-Parallel tests 是擁有相同 TIF 的多份測驗; 而 TCC-parallel tests，是擁有相同測驗特性的

多份測驗[23]。當平行測驗同時符合 TCC 和 TSIF，就可提升測驗間等化(比較)的精確度。然而 TCC/TSIF-parallel tests 除了要有相同的 TIF 之外，測驗特性(像是測驗的試題數、內容、題型和長度)也要完全相同。由於符合 TCC/TSIF 的平行測驗組裝難度很高，過去的測驗組裝研究較少著墨於這個議題。因此，本文提出一改良型的株落選擇演算法(CLONAL)來同時組裝 TCC/TSIF-parallel tests。藉由 CLONAL 的平行搜尋機制和獨立發展特性來解決同時組裝的變數和限制式過多問題，使得該方法不需用串接測驗的方式，也不需以一份大型試題集合來代替多份測驗的作法，便能以單一組裝模型來同時組裝多份平行測驗。且其強大的搜尋能力使得組裝測驗能進一步的滿足 TCC/TSIF-parallel 的嚴苛測驗規格。

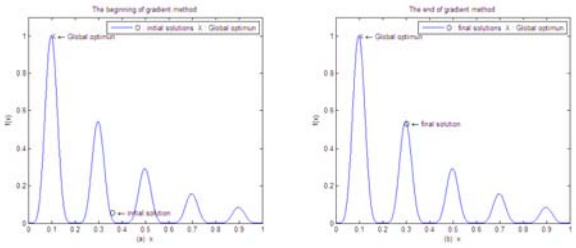
本文的架構統整如下，在第二章介紹株落選擇演算法的起源、特性和核心步驟。在第三章介紹本研究所考慮的兩種平行測驗組裝問題模型。在第四章介紹用來解決平行測驗組裝問題的改良型株落選擇演算法的詳細步驟和改良要點。第五章將介紹實驗方式以及呈現一系列的實驗結果，最後一章針對本文作一個分析和討論。

2.株落選擇演算法

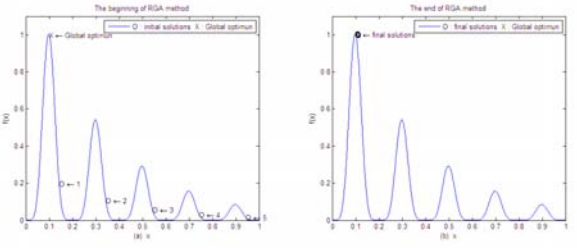
株落選擇演算法是 De Castro & Von Zuben 提出的啟發式演算法[24]。該演算法是基於生物免疫系統中的株落選擇法則的概念設計而成，特點是能同時找出搜尋空間中的多組解。先介紹株落選擇演算法中的平行搜尋機制和獨立發展特性，以了解為何 CLONAL 非常適合用來解決平行測驗組裝問題。以一典型的尋找最大值問題為例，從搜尋方式來了解不同方法間的差異:

傳統的搜尋方法：例如梯度法採用單點搜尋的方式，可以在很短的時間內找到最佳解。然而，當搜尋點位於局部最佳解時，便會因周圍的解都比局部最佳解小，導致難以跳脫出來而失去找到全域最佳解的機會。如圖三，當初始解不是落於最佳解附近時，就很容易陷入相對較近的局部最佳解。

基因演算法(Genetic algorithm, GA)、粒子群最佳化演算法：採用平行搜尋機制，多點同時搜尋可避免陷入局部最佳解，也增加了找到全域最佳解的可能性。然而，由於GA的所有搜尋點會朝最佳解的方向移動，所以最後找出的多個解會非常相似。如圖四，在初期時的初始解散佈在搜尋空間中，但到了後期所有解都會集中在最佳解附近。因此無法在一次搜尋中找出多個不同的解，必須進行多次搜尋才能解決平行測驗組裝問題(循序組裝)。

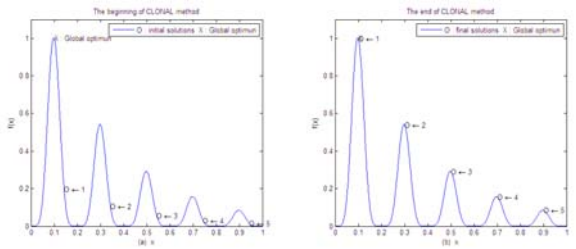


圖三. 梯度法的搜尋策略



圖四. 基因演算法的搜尋策略

株落選擇演算法:CLONAL和基因演算法一樣採用平行搜尋，不同的是族群中的每個解都是獨立發展，不會朝族群中最佳解的方向移動。如圖五，在初期時所有初始解和GA一樣分散在搜尋空間中。不同的是，CLONAL到了後期所有解依然分散在搜尋空間中，並各自找出了所在區域中的局部最佳解。由於CLONAL採用獨立發展的方式保有了族群的多樣性，因此很適合用來找出搜尋空間中的多個解。而這正是平行測驗組裝問題的需求。不但可以避免循序組裝產生的不平等問題，且只要用組裝單一測驗的問題模型就可以同時組裝多份測驗，不需引入額外的選擇變數和限制條件。接下來將介紹株落選擇演算法的起源和核心概念。



圖五. 株落選擇演算法的搜尋策略

2.1 株落選擇原則

CLONAL是根據免疫系統的概念設計而成。生物的免疫系統主要包含了兩道防衛線，一種是先天免疫系統，另一種是適應性免疫系統。而適應性免疫系統的核心就是株落選擇法則。當B淋巴細胞遇到抗原，抗原會活化B淋巴細胞產生抗體分子。由於抗體分子附屬於B淋巴細胞，因此在此將兩者視為同一個體。每個B淋巴細胞能夠產生唯一一種獨特的抗體分子，該抗體能夠識別出特定的抗原並和其結合。而這種結合的行為，將刺激B淋巴細胞複製(增殖)出一個細胞或多個細胞，這些被複製而成的細胞就稱為克隆(clone)。而B淋巴細胞與抗原的匹配程度便稱為親和力。在增殖的過程中，B淋巴細胞上的抗體會有少量的突變，突變可能會提升部分抗體的親和力。這些擁有更高親和力的抗體將有更多得機會進行克隆選擇。因此具有高親和力的克隆群體將越來越龐大，與抗原的親和力也會越來越高。

簡單來說，株落選擇原則的關鍵在於選出高親和力的抗體，以及將這些高親和力的抗體大量增殖再進行細微的突變，使抗體的親和力能不斷提升。CLONAL 的核心便是由這三個操作所組裝的。回應到測驗組裝的問題，每個抗體就是代表一份測驗，選出較優良的測驗進行些微修正(突變)，以更滿足於目標測驗的規格(抗原)，不斷重複這樣的歷程來產生最匹配測驗規格(親和力最高)的試題組合。接下來將詳細介紹基於自免疫系統概念所發展而成的株落選擇演算法流程。

2.2 株落選擇演算法流程

株落選擇演算法中主要有兩種機制：擇優和突變。這兩種機制是免疫系統特性的實踐。演算法的方塊圖(圖六)及主要步驟的描述如下：

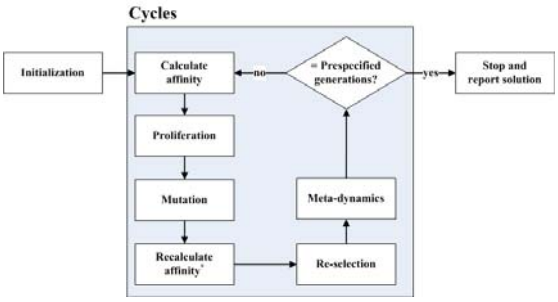
主要步驟

1. 初始(initialization)：創造一個隨機的抗體族群，此族群中包含*popu_size*個抗體，每一個抗體即代表一份測驗
2. 計算親和力(calculate affinity)：親和力表達的是抗體對抗原的適配程度。在測驗組裝問題中，親和力代表組裝測驗與測驗規格之間的差距。差距越小表示親和力越高。而測驗規格將在之後介紹測驗組裝問題的模型時，以方程式的方式表達。
3. 株落選擇：選出N_c個最高親和力的抗體來進行株落選擇，對每個被選出的抗體執行以下步驟

   3.1 增殖(proliferation)：將抗體增殖來組裝克隆群體，群體中的抗體數(克隆數)與原染色體的親和力成正比

   3.2 突變(mutation)：對每個克隆群體進行突變，被突變的抗體數由突變率決定，而突變率和抗體的親和力成反比。

   3.3 重選(re-selection)：重新計算出克隆族群中所有抗體的親和力，然後選擇最高親和力的抗體回族群取代原抗體
4. 後期變動(meta-dynamic)：從族群中選擇*popu_size* - N_c個最低親和力的抗體，用全新的隨機抗體取代掉
5. 循環(cycle)：重複步驟2到4，直到完成指定的世代數



圖六. 株落選擇演算法的演算法流程圖

介紹完株落選擇演算法的流程，可知用CLONAL來解決測驗組裝問題的關鍵在於抗體(測驗)的表達和親和力的計算。抗體表達已在先前的部份介紹過了(見圖二)，而親和力計算函數是針對測驗組裝問題模型的目標函數來設計的。所以接下來將介紹本文所要考慮的測驗組裝問題模型。

3. 測驗組裝問題模型

在本文中考慮了以固定訊息量為目標的測驗組裝問題模型。針對CLONAL本身的特性，本文對原先的問題模型進行兩點調整:

1. 選擇變數為整數型態

   一般來說題庫試題數>>測驗長度，而二進制的試題選擇變數數量等於題庫試題數，導致測驗組裝問題需要非常大量的選擇變數，在使用平行搜尋機制的演算法中尤其嚴重。本文的試題選擇變數為整數型態(試

題序號)，選擇變數數量等於測驗長度。這樣不但能降低選擇變數的數量，且不需要測驗長度的限制式(3)，使得問題的模型得以簡化。

2. 去掉平行測驗數F

藉由演算法的機制來避免測驗間發生試題重疊，且每份測驗(抗體)是分別進行組裝。所以只需考慮單一抗體的親和力計算，不需要加入平行測驗數F和試題重疊的限制式(4)。因此，本文得以將同時組裝模型大大簡化成單一測驗組裝的問題模型，而不需引入的選題變數和限制式。

另外，本文所考慮的評估準則包括內容、解題技巧、題型和作答時間。為了要組裝符合TCC/TSIF的平行測驗，必須要用目標值來設定測驗規格(如圖七b)。過去受限於組裝技術的能力，通常將測驗規格設定成一組較寬鬆的目標範圍(如圖七a)。然而，這樣的方式所產生的平行測驗不可能擁有相同的測驗規格，無法滿足TCC/TSIF-parallel。因此，本文用目標值作為測驗規格來組裝完全相同的平行測驗，即使組裝問題的難度會因此提升很多。

將 $\theta=0$ 所在的測驗訊息量最大化 S.t. 第一章至少要有 5 題 第一章不要超過 30 題 第二章至少要有 20 題 選擇題至少要有 25 題 是非題不要超過 10 題 測驗長度為 40 題 總字數不要超過 1800 字

(a)以目標範圍設定的測驗規格

將 $\theta=0$ 所在的測驗訊息量最大化 S.t. 第一章要有 15 題 第二章要有 25 題 選擇題要有 35 題 是非題要有 5 題 測驗長度為 40 題 總字數不要超過 1800 字

(b) 以目標值設定的測驗規格

圖七. 測驗規格

3.1 固定訊息量問題模型

在此模型中，施測者依照測驗的類型和需求，直接指定目標TIF的數值。因此模型的目標就是要滿足固定的TIF目標$T(\theta_k)$。

- x_i:實數型態的選擇變數，其值代表題庫中的試題序號
- $c_{x_i h}$:二進制的內容屬性變數，當$c_{x_i h}$=1時，表示試題x_i屬於內容h，反之則否
- C_h:內容屬性h的目標試題數
- $s_{x_i v}$:二進制的解題技巧屬性變數，當$s_{x_i v}$=1時，表示試題x_i屬於內容v，反之則否
- S_v:解題技巧屬性v的目標試題數
- $t_{x_i m}$:二進制的題型屬性變數，當$t_{x_i m}$=1時，表示試題x_i屬於內容m，反之則否
- T_m:題型屬性m的目標試題數
- r_{x_i}:試題x_i的作答時間，
- $R^{(u)}$:目標測驗作答時間

$$\text{Minimize} \quad \sum_{k=1}^{K}\sum_{i=1}^{n} I_{x_i}(\theta_k) \tag{10}$$

S. t.:

$$\sum_{i=1}^{n} I_{x_i}(\theta_k) - T(\theta_k) \geq 0, k = 1,...,K \tag{11}$$

$$\sum_{i=1}^{n} c_{x_i h} = C_h, h = 1,...,H \tag{12}$$

$$\sum_{i=1}^{n} s_{x_i v} = S_v, v = 1,...,V \tag{13}$$

$$\sum_{i=1}^{n} t_{x_i m} = T_m, m = 1,...,M \tag{14}$$

$$\sum_{i=1}^{n} r_{x_i} \leq R^{(u)} \tag{15}$$

$$x_i \in [1, N], i = 1,...,n \tag{16}$$

限制式(11)和目標函數(10)與限制式(2)和目標函數(1)相同，只是少了平行測驗數F。而限制式(12)、(13)、(14)分別定出了內容、解題技巧和題型的目標試題數。限制式(15)是指出組裝測驗的總字數不得超過上限。介紹完組裝模型以及避免試題重疊的作法後，接下來將詳細說明如何用改良型株落選擇演算法來解決此模型。

3.2 改良型株落選擇演算法流程

本文的關鍵在於改良CLONAL的演化機制來避免測驗(抗體)間產生試題重疊。在CLONAL中，有三個階段會引入新的試題導致

試題重疊，分別是初始、突變和後期變動。為了確保在這三階段中所引入的新試題是其他測驗所沒有的，本文加入了封鎖試題(block item)的概念。試題一旦被列為封鎖試題，就會暫時從題庫中剔除，也就不會有機會被其他測驗所選擇。因此，在初始階段時，每產生一份隨機測驗，就把該測驗中所有試題列為封鎖試題，這樣接下來產生的隨機測驗就不可能與先前測驗中的試題重疊。同樣的道理，在突變階段時，所有測驗的試題都已列為封鎖試題，因此產生的新試題也不會與族群中的試題重疊。在後期變動的階段也是如此。用這樣的作法，可使每份測驗(抗體)所選的試題都不會重複，因此可輕鬆的同時組裝多份平行測驗，平行測驗組裝問題的複雜度也不會隨著測驗數F的增加而提升。以下介紹用來解決固定訊息量問題的改良型株落選擇演算法

1. 初始：

用隨機的方式依序產生*popu_size*個抗體(測驗)來組裝族群，每個抗體x為n個實數組裝的字串，實數值介於 1~N(見圖二)。逐次的隨機產生抗體，在產生的同時其所選擇的試題立刻列入禁忌列表中，以避免下一個抗體選到相同的試題。

2. 計算親和力：

計算族群中每一個抗體的親和力。親和力代表了此測驗滿足測驗規格的程度，因此要依據模型中的目標函數和所有限制式來設計親和力函數。在固定訊息量的模型中，限制式主要目的都是定出目標值，因此親和力函數是要將組裝測驗與目標值的偏差轉換成對親和力的懲罰值。

針對限制式(11)，以$D^{(\mathrm{I})}$表示組裝測驗在所有能力等級的訊息量總偏差值

$$\sum_{k=1}^{K}\left|\left[\sum_{i=1}^{n} I_{x_i}(\theta_k)\right]-T(\theta_k)\right| = D^{(\mathrm{I})} \qquad (17)$$

針對限制式(12)，以$D^{(\mathrm{C})}$表示組裝測驗在所有內容屬性的總偏差試題數

$$\sum_{h=1}^{H}\left|\left(\sum_{i=1}^{n} c_{x_i h}\right)-C_h\right| = D^{(\mathrm{C})} \qquad (18)$$

針對限制式(13)，以$D^{(\mathrm{S})}$表示組裝測驗在所有解題技巧屬性的總偏差試題數.

$$\sum_{v=1}^{V}\left|\left(\sum_{i=1}^{n} s_{x_i v}\right)-S_v\right| = D^{(\mathrm{S})} \qquad (19)$$

針對限制式(14)，以$D^{(\mathrm{T})}$表示組裝測驗在所有題型屬性的總偏差試題數

$$\sum_{m=1}^{M}\left|\left(\sum_{i=1}^{n} t_{x_i m}\right)-T_m\right| = D^{(\mathrm{T})} \qquad (20)$$

針對限制式(15)，以$D^{(\mathrm{R})}$表示組裝測驗總字數與目標上限的偏差值

$$\max\left\{0,\left(\sum_{i=1}^{n} r_{x_i}\right)-R^{(\mathrm{u})}\right\} = D^{(\mathrm{R})} \qquad (21)$$

針對目標式(10)的親和力計算函數

$$\begin{aligned}\mathrm{affinity}(x) = {} & W^{(\mathrm{I})}D^{(\mathrm{I})}+W^{(\mathrm{C})}D^{(\mathrm{C})}+ \\ & W^{(\mathrm{S})}D^{(\mathrm{S})}+W^{(\mathrm{T})}D^{(\mathrm{T})}+W^{(\mathrm{R})}D^{(\mathrm{R})}\end{aligned} \qquad (22)$$

當所有偏差值都等於 0 時，組裝測驗就完全符合目標測驗了。因此(22)式的目標是將所有評估準則的偏差值權重和作最小化，其中各種權重W是使用者自訂的正規化因子，用來放大和平衡不同限制式之間的偏差值。

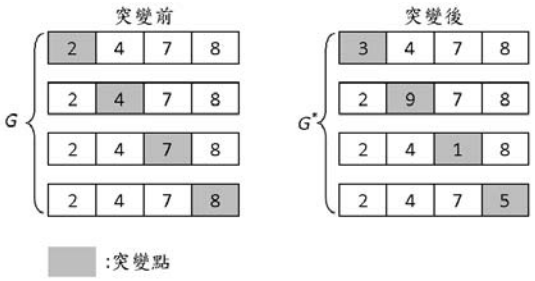
3. 株落選擇：

計算完適應值後，選出N_c個親和力最高的抗體進行株落選擇。本文針對組裝問題中固定測驗長度的特性，將株落選擇進行改良，以進行密集的小範圍搜尋。對每個要進行株落選擇的抗體x進行以下步驟:

3.1 增殖：在原本CLONAL中，克隆的數量與親和力成正比。但在本文中，克隆的數量直接設定成n，因此每個克隆群組G中包含了n個抗體。

3.2 單點突變：突變的數量也不是由突變率決定，而是直接對群體G中的每一個抗體進行單點突變。作法是依序將第m個抗體所選擇的第m題替換(突變)成其它不重複選擇的試題，突變後的群體稱為G^*。以圖八為例，測驗的長度為 4，因此克隆群組G中有 4 個抗體x的複製。在步進突變時，直接選擇第一個x的第一個基因為突變點，經隨機選擇選出一道新的試題來替換掉第一個基因。為了避免選到x已經選擇的試題，必須將x原有的試題列為封鎖試題，在此例中的封鎖試題為(2、4、7、8)。避開封鎖試題後，經隨機選出的第 3 題被用來替換掉第 2 題，並把第 3 題加入封鎖試題，這樣就完成了x的突變；再選擇第二個x的第 2 個基因為突變點，經隨機選擇後替換成第 9 題；其餘同理。上述作法是針對產生單一測驗的單點突變。由

於本文目的是要同時產生多份試題不重疊的平行測驗，因此封鎖的就不只是x所選擇的試題，而是整個族群中所有抗體所選擇的試題，這樣就可確保在突變程序中不會發生試題重疊的現象。



圖八. 株落選擇中的單點突變範例

3.3 重選：重新計算突變後的群體 G^* 中所有抗體的親和力，挑出最高親和力的抗體 x^*，並比較 x 和 x^* 的適應值，若 $x^*>x$，則以 x^* 取代掉原族群中的 x；反之則保留 x。

4. 後期變動：

從族群中選擇 $popu\_size - N_c$ 個最低親和力的抗體，用全新的隨機抗體取代掉。在產生隨機抗體的程序時也排除了整個族群中所選擇的試題，因此不會發生試題重疊的現象。

5. 循環：

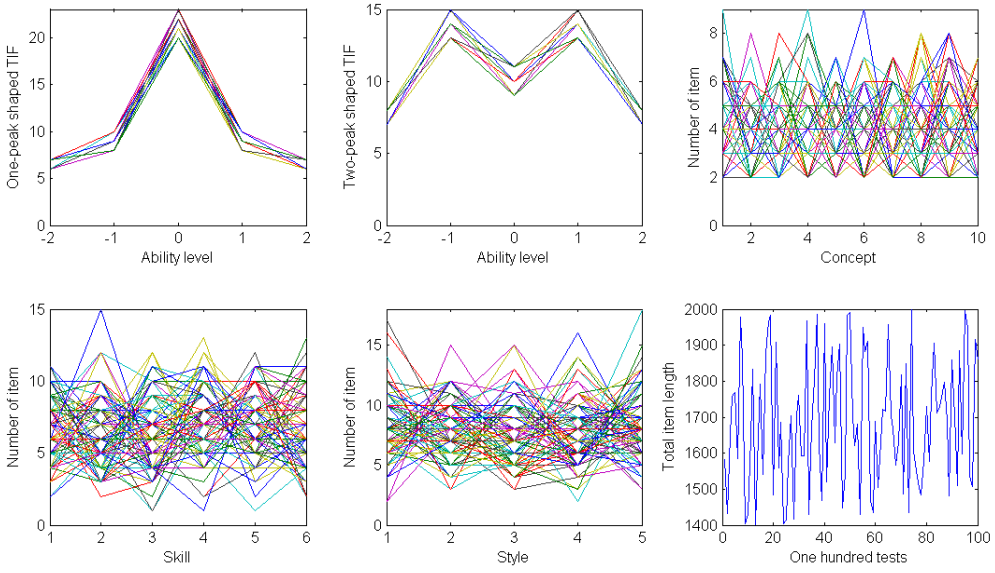
重複步驟 2 到 4，直到完成指定的世代數 N_{gen}，然後從族群中挑出 F 個親和力最高的抗體作為最終解。這些抗體就是 F 份測驗規格相同，但試題不重疊的 TCC/TSIF Parallel tests。

4. 實驗設計與結果

在本部分呈現CLONAL的組裝實驗結果。為了進行大量的組裝實驗，本文參考了Sun的實驗設計來建置一個4000題的虛擬題庫[13]。表一顯示每道試題參數的隨機生成範圍，包含試題的三參數和屬性值。建立在此虛擬題庫上，便可設定各式各樣的測驗規格來組裝目標測驗，以進行演算法的效能評估實驗。接下來將針對固定訊息量的問題模型來進行實驗。

本文中考慮兩種TIF波形，分別是單峰型和雙峰型的TIF。為了評估CLOANL是否能滿足各式各樣的測驗規格，將分別對這兩種類型的TIF各進行一百次的組裝，每次組裝的目標TIF都是隨機產生。對單峰型TIF，目標訊息量的隨機目標值範圍為6~7, 8~10, 20~23, 8~10, 6~7，這些範圍分別對應到能力等級θ= –2, –1, 0, 1, 2; 而雙峰型TIF的隨機目標值範圍為7~8, 13~15, 9~11, 13~15, 7~8。依照這些範圍，分別為兩種類型的TIF產生100種目標TIF。而內容、技巧、題型和測驗長度的目標試題數也都是隨機產生。圖九即為一百次隨機測驗規格的數值分布圖，每一條線即代表一次測驗組裝的評估準則。從圖中可以看出，本實驗所考慮的測驗規格非常多樣。

為了要測試CLONAL用於測驗組裝的效能，將與之前的方法作比較。首先採用Premium



圖九. 百次隨機測驗規格的目標值分佈圖

表一. 模擬題庫的屬性

Attributes Information	三參數			試題屬性			
	鑑別度	困難度	猜測度	內容	技能	題型	長度
範圍	0.8~3.0	-3.0~3.0	0.1~0.3	1~10	1~6	1~5	20~50
類型	real	real	real	integer	integer	integer	integer
平均值	1.9	-0.05	0.2	5.444	3.494	2.967	34.854
標準差	0.638	1.734	0.058	2.892	1.715	1.429	10.149

表二. RGA和CLONAL的演算法參數設定

RGA			CLONAL		
N	4000	題庫題數	N	4000	題庫題數
n	40	測驗試題數	n	40	測驗試題數
popu_size	100	族群數	*popu_size*	30	族群數
P_c	1	交配率	N_c	20	株落選擇數
P_m	1/n	突變率	*clone_size*	5	平行測驗數
N_{gen}	1000	世代數	N_{gen}	1500	世代數

Solver Platform的the Large-Scale LP/QP Solver Engine來進行LP方法的組裝實驗[25]。然而，由於本文考慮的評估準則非常嚴苛，結果顯示用LP組裝單份試卷需花一周的時間，幾乎無法在合理時間內同時組裝多份TCC/TSIF Parallel tests，因此LP無法作為比較對象。另外，與CLONAL同屬於啟發式演算法的GA已被證實能有效的解決較複雜的測驗組裝問題[9, 13-14, 20]，因此本文將用GA與CLONAL進行比較。對於GA的原理及使用方式，先前文獻已介紹的非常詳細，本文就不在此贅述。由於GA本質上不適合用來同時組裝多份測驗，因此將以循序組裝的方式，與CLONAL同時組裝5份測驗進行比較。表二呈現兩種演算法的參數設定。此外，訊息量偏差值的精確率設定在小數點後4位數，因此W=10000。

分別用RGA和CLOANL進行一百次測驗組裝，每次記錄下組成測驗與$T(\theta_k)$的均方差值，再將其百次的均方差取平均後呈現在表三。實驗結果指出RGA和CLOANL都能有效的組裝擁有相同測驗特性的測驗(滿足TCC-Parallel)。然而，CLOANL不但可以同時組裝5份試卷，且能大大降低與目標TIF的偏差值。在單峰和雙峰，CLOANL的平均偏差分別比RGA小8.71和8.28倍。

表三. RGA和CLONAL比絕對訊息量

MSE	絕對訊息量			
	單峰型 TIF		雙峰型 TIF	
	RGA	CLONAL	RGA	CLONAL
1st	0.68	0.04	0.51	0.03
2st	0.64	0.05	0.73	0.05
3st	0.62	0.07	0.64	0.07
4st	0.51	0.09	0.53	0.08
5st	0.60	0.11	0.52	0.09
Ave.	0.61	0.07	0.58	0.07
Ratio	8.71	1	8.28	1

5. 結論

啟發式演算法已被證實比 LP 更適合用來解決平行測驗的組裝問題[20]，在本文中提出一個改良式的 CLONAL 來同時組裝多份 TCC/TSIF-parallel tests。該方法能將同時組裝多份測驗的問題簡化成單一測驗組裝問題，因此完全解決了同時測驗組裝問題中，大量的選擇變數和限制式造成測驗品質下降的問題。經實驗結果得知，在題庫有足夠試題的情況下，該方法能有效的同時組裝多份滿足目標測驗規格的多份平行測驗。與 RGA 以循序方式組裝多份平行測驗相比，CLONAL 不但能同時組裝多份 parallel tests，且有較小的偏差。整體來說，使用 CLONAL 來組裝測驗有以下優點

1. 能同時組裝多份測驗
   不會發生循序組裝的不平等問題，也不會因平行測驗數的提升，而大幅增加問題的複雜度
2. 求解能力強
   即便是如 TCC/TSIF-parallel tests 的嚴苛組

裝條件，也能在可接受的時間內組裝符合目標測驗規格的試題組合。不會因試題數或限制式的增加，就無法組裝目標測驗。且題庫的試題數越多，越能突顯啟發式演算法的優勢。

3. 擴充性強
   若要加入新的限制式，只要修改適應值計算式即可，不需更動到演算法的搜尋程序
4. 符合 Belov 所提出的均勻組裝概念[8]，由於 CLONAL 隨機從題庫挑選試題，可避免少數優質試題被重複挑選，曝光率過高導致測驗信度下降。

基於上述優點，BGA*-C 可以被適當地運用測驗系統中，用來評估學生的學習狀態，並在今後的工作中實作於真實的評量系統。

誌謝

非常感謝國科會計畫編號：NSC98-2622-E-018-004-CC3 之贊助。

參考文獻

[1] van der Linden, W. J., "Optimal Assembly of Psychological and Educational Tests," ***Applied Psychological Measurement,*** vol. 22, pp. 195-211, 1998.

[2] Hambleton, R. K. and Swaminathan, H., ***Item Response Theory: Principles and Applications.*** Netherlands: Kluwer Academic Publishers Group, 1985.

[3] Ackerman, T. A., "An alternative methodology for creating parallel test forms using the IRT information function," in ***The Annual Meeting of the National Council for Measurement in Education***, 1989.

[4] Adema, J. J., **"**Methods and Models for the Construction of Weakly Parallel Tests," ***Applied Psychological Measurement,*** vol. 16, pp. 53-63, 1992.

[5] Armstrong, R. D. and Jones, D. H., "Polynomial Algorithms for Item Matching," ***Applied Psychological Measurement,*** vol. 16, pp. 365-371, 1992.

[6] Armstrong, R. D., Jones, D. H., Li, X., and Wu, I. L., "A Study of a Network-Flow Algorithm and a Noncorrecting Algorithm for Test Assembly," ***Applied Psychological Measurement,*** vol. 20, pp. 89-98, 1996.

[7] Sun, K. T. and Chen, S. F., **"**A study of applying the artificial intelligent technique to select test items," ***Psychological Testing,*** vol. 46, pp. 75-88, 1999.

[8] Belov, D. I. and Armstrong, R. D., "Monte Carlo Test Assembly for Item Pool Analysis and Extension," ***Applied Psychological Measurement,*** vol. 29, pp. 239-261, 2005.

[9] Hwang, G. J., Lin, B. M. T., Tseng, H. H., and Lin, T. L., "On the development of a computer-assisted testing system with genetic test sheet-generating approach," ***IEEE Transactions on Systems, Man, and Cybernetics-Part C: Applications and Reviews,*** vol. 35, pp. 590-594, 2005.

[10] Hwang, G. J., Yin, P. Y., and Yeh, S. H., "A tabu search approach to generating test sheets for multiple assessment criteria," ***IEEE Transactions on Education,*** vol. 49, pp. 88-97, 2006.

[11] Yin, P. Y., Chang, K. C., Hwang, G. J., Hwang, G. H., and Chan, Y., "A Particle Swarm Optimization Method to Composing Serial Test-sheets for Multiple Assessment Criteria," ***Educational Technology and Society,*** vol. 9, pp. 3-15, 2006.

[12] Hwang, G. J., Chu, H. C., Yin, P. Y., and Lin, J. Y., "An innovative parallel test sheet composition approach to meet multiple assessment criteria for national tests," ***Computers & Education,*** vol. 51, pp. 1058-1072, 2008.

[13] Sun, K. T., Chen, Y. J., Tsai, S. Y., and Cheng, C. F., "Creating IRT-Based Parallel Test Forms Using the Genetic Algorithm Method," ***Applied Measurement in Education,*** vol. 21, pp. 141 - 161, 2008.

[14] Finkelman, M., Kim, W., and Roussos, L. A., "Automated Test Assembly for Cognitive Diagnosis Models Using a Genetic Algorithm," ***Journal of Educational Measurement,*** vol. 46, pp. 273-292, 2009.

[15] van der Linden, W. J. and Adema, J. J., "Simultaneous Assembly of Multiple Test Forms," ***Journal of Educational Measurement,*** vol. 35, pp. 185-198, 1998.

[16] van der Linden, W. J., ***Linear Models for Optimal Test Design***. Berlin: Springer, 2005.

[17] Theunissen, T. J. J. M., "Binary programming and test design," ***Psychometrika,*** vol. 50, pp. 411-420, 1985.

[18] Boekkooi-Timminga, E., "The Construction of Parallel Tests From IRT-Based Item

Banks," ***Journal of Educational and Behavioral Statistics,*** vol. 15, pp. 129-145, 1990.

[19] Boekkooi-Timminga, E., "Simultaneous test construction by zero-one programming," ***Methodika,*** vol. 1, pp. 101-112, 1987.

[20] Verschoor, A. J., "Genetic algorithms for automated test assembly," ***Enschede***, 2007.

[21] Lee, C. L., Huang, C. H., and Lin, C. J., "Test-Sheet Composition Using Immune Algorithm for E-Learning Application," in ***New Trends in Applied Artificial Intelligence***, ed, 2007, pp. 823-833.

[22] Breithaupt, K., Ariel, A., and Veldkamp, B. P., "Automated Simultaneous Assembly for Multistage Testing," ***International Journal of Testing,*** vol. 5, pp. 319 - 330, 2005.

[23] McDonald, R. P., ***Test theory : a unified treatment***. Mahwah, N.J.: L. Erlbaum Associates, 1999.

[24] de Castro, L. N. and Von Zuben, F. J., "Learning and optimization using the clonal selection principle," ***IEEE Transactions on Evolutionary Computation,*** vol. 6, pp. 239-251, 2002.

[25] Cor, K., Alves, C., and Gierl, M. J., "Conducting Automated Test Assembly Using the Premium Solver Platform Version 7.0 With Microsoft Excel and the Large-Scale LP/QP Solver Engine Add-In," ***Applied Psychological Measurement,*** pp. 652-663, 2008.