試題反應理論在教育測驗上之應用

郭伯臣[1]、吳慧珉[2]、陳俊華[3]
[1、3]台中教育大學教育測驗統計研究所
[2]國家教育研究院測驗及評量研究中心

摘要

E本文主要簡介試題反應理論的發展與應用，首先，比較試題反應理論與古典測驗理論之差異，接著介紹目前常用試題反應理論模式，例如：單向度單參數、二參數、三參數Logistic模式及多向度模式，進而簡介新發展之高階試題反應理論模式並舉例說明使用上之差異，最後，說明試題反應理論在大型教育測驗及電腦化適性測驗上之應用。

Application of IRT in Educational Measurement

Bor-Chen Kuo[1]、Huey-Min Wu[2]、Chun-Hua Chen[3]
[1、3]Graduate Institute of Educational Measurement and Statistics National Taichung University of Education
[2]Research Center for Testing and Assessment, National Academy for Educational Research

Abstract

EThe goal of this paper is to introduce the development and application of item response theory. First, the differences of item response theory and classical test theory are compared. Second, some useful and poplar item response models are mentioned, such as unidimensional one-parameter, two-parameter, three-parameter logistic models and multidimenstional models. Moreover, some recently proposed higer-order item response models are also described. Final part contains the applications of item response theory to the large scale education assessments and computerized adaptive tests.

十九世紀初比奈賽門(Bi-net Simon)智力量表問世，測驗理論開始受到學者重視，其中廣爲人知的應屬古典測驗理論。古典測驗理論(classical test theory, CTT)最早是Gulliksen於1950出版的書—心理測驗理論(Theory of Mental Test)被介紹，古典測驗理論的模式簡單易懂成爲二十世紀測驗理論的主軸，在測驗編製、評估及使用上有其貢獻，然隨著測驗需求量的日益增加及形式多樣化，其簡單的假設造成古典測驗理論應用之限制，其中最明顯的限制是不同版本測驗的比較問題，即不同的學生參加不同的測驗，如何比較這些學生的成績，於是不同的測驗理論陸續被提出，如Lord 和Novick(1968)提出以模式爲基礎的測驗理論，即現在常聽到試題反應理論(item response theory, IRT)的先驅，但由於其理論模式過於艱澀且估計過程繁雜，再加上當時的電腦設備並不先進，故此書並未引起太大的共鳴。Lord(1980)出版「試題反應理論於測驗的應」一書，正式介紹試題反應理論，由於試題反應理論架構嚴謹，考慮層面廣泛，除了延續古典測驗理論的功能，並藉由電腦科技協助，突破古典測驗理論在應用上的瓶頸，此時電腦技術已較進步，成爲當前的主流測驗理論之一。本文主要介紹試題反應理論與其在教育測驗之應用，主要內容包含試題反應理論與古典測驗理論之差異、試題反應理論模式簡介、試題反應理論之應用等。

一、試題反應理論與古典測驗理論之差異

古典測驗理論也被稱爲真實分數理論(true score theory)，主要是建立於簡單的線性函數假設，$X = T + E$ ，其中 X 爲可觀察分數，T 爲真實分數，E 爲誤差分數(Lord & Novick, 1968)。古典測驗理論主要考量整份測驗的總分解釋學生的能力。而試題反應理論是以非線性數學模式爲基礎，並以試題的觀點解釋學生的能力，學生在某一題試題的表現，如答對或答錯，和學生所具備的某一種能力，如數學能力，具有一種非線性關係，這一種非線性關係可透過一條連續性遞增的數學函數表示，稱爲試題特徵曲線(item characteristic cure, ICC)。

古典測驗理論與試題反應理論對於測量問題所抱持的觀點並不一樣，試題反應理論具備幾項特點，正好可以補足古典測驗理論之限制：

1. 在誤差假設方面，在古典測驗理論中是假設所有的受試者都有相同的測量標準誤；試題反應理論中假設每位具有不同能力水準的受試者，對應不同的測量標準誤，如果使用的是一份難度適中的測驗，理論上對於中等能力的受試者會有較小的測量標準誤，而對於較高能力或能力差的受試者，其測量標準誤會較大。

2. 在信度方面，相較於短測驗，古典測驗理論假設測驗題數越多的測驗其信度指數值亦較高；試題反應理論則假設題數少的測驗亦可以得到很好的信度指數，如適性測驗即是最好的例子。
3. 在試題參數估計方面，古典測驗理論中，試題的難度指數和鑑別度指數，會因爲受試者的能力分配不同而得到不一樣的估計結果，如要得到不偏的估計結果，必須選取具有代表性的樣本；而試題反應理論具有參數不變性(parameter invariance)之特色，即試題的試題參數估計，不受受試者能力分布影響，受試者能力值估計，亦不受到測驗試題之影響(Hambleton & Swaminathan, 1985)。
4. 在分數解釋方面，古典測驗理論是透過常模參照的方式解釋分數的意義，受試者只知道他贏過誰，卻無法得知會什麼；試題反應理論則是將能力與試題擺在同一個量尺，受試者可以透過能力與試題難度的差異，瞭解自己大概可以答對哪些題目。
5. 在題型設計方面，對於古典測驗理論，混合題型的設計，如一份測驗同時有多元計分和二元計分會對總分的計算產生偏誤；然而這一種混合題型的設計在試題反應理論卻能得到最佳的估計效果。
6. 在作答反應組型解釋方面，受試者能力的估計主要依賴於受試者的作答反應組型和所施測的試題特性，以古典測驗理論而言，不管受試者作答反應組型如何，只要加總之總分相同，即代帶能力相同；而對於試題反應理論，原始總分相同的受試者，若是其作答反應組型不一樣，亦有可能得到不同的能力估計值。
7. 在等化方面，古典測驗理論假設唯有複本測驗(parallel test)，不同的測驗分數才能進行比較，且結果是最佳的，試題反應理論則無此假設，而等化效果最佳的情況是不同測驗所使用的題目能涵蓋不同能力學生的需求。

二、試題反應理論模式簡介

教育現場，教師發展一份測驗時必須先決定測量的目標，也就是所欲測量的學生能力是什麼？這個目標可以是單一的，如界定義爲數學能力，亦可以是多維的，如欲測量學生之幾何能力、計算能力等，而後展開命題，最常使用的測驗題型是選擇題、填充題和應用問題等題型，針對不同的題型會搭配不同的計分方式，如選擇題型和填充題型可使用答錯0分，答對1分，這種計分模式在測驗理論中稱爲「二元(dichotomous)計分」，而應用問題則可以是答錯0分、部分答對1分、全對2分，這種計分模式則被稱爲「多點(polytomous)計分」。

試題反應理論假設學生所具備的某一種能力和受試者在某一題的答題情形，可以透過數學模式，也就是試題特徵曲線(ICC)加以闡述。依據對於能力特質之基本假設不同和計分模式的不同，主要可分爲單向度試題反應理論(unidimensional IRT, UIRT)、多向度試題反應理論(multidimensional IRT, MIRT)以及；依據計分型態，可分爲二元(dichotomous)計分和多點(polytomous)計分模式，目前更有學者結合單向度試題反應理論與多向度試題反應理論，發展高階層試題反應理論(higher-order IRT, HO-IRT)。以下將介紹這幾種理論模式：

（一）單向度試題反應理論模式

單向度試題反應理論模式必須符合單向度(unidimensionality)、局部獨立(local independence)、非速度性(nonspeedness)、「知道-正確」假設(“know-correct” assumption)四項基本的假設(Weiss & Yoes, 1991)，才能進行測驗資料之分析。

基於試題反應理論的單向性假設，一般使用之試題反應理論爲單向度試題反應理論，本研究介紹二元計分對數型模式及多點計分模式，二元計分對數型模式包含有單參數對數模式(one-parameter logistic model, 1PL)、二參數對數模式(two-parameter logistic model, 2PL)及三參數對數模式(three-parameter logistic model, 3PL)；多點計分模式包含部分給分模式(partial credit model, PCM)和廣義部分給分模式(generalized partial credit model, GPCM)。

1. 二元計分模式

(1)單參數對數模式

單參數對數模式有Rasch模式之稱(Rasch, 1960; Wright & Stone, 1979; Wright & Master, 1982)，在試題反應理論的1PL模式下，假設受試者 j 之能力爲 θ_j，其作答試題 i 通過的機率如下：

$$P(X_{ij}=1\,|\,\theta_j, b_i)=\frac{1}{1+\exp[-(\theta_j-b_i)]} \qquad \text{（公式1）}$$

其中，X_{ij} 爲受試者 j 在試題 i 的作答反應，答對記爲1，答錯記爲0；b_i 爲試題 i 之試題難度參數(item difficulty parameter)。

(2)二參數對數模式

在試題反應理論的2PL模式下，假設受試者 j 之能力爲 θ_j，其作答試題 i 通過的機率如下(Birnbaum, 1968)：

$$P(X_{ij}=1|\theta_j,b_i,a_i)=\frac{1}{1+\exp[-a_i(\theta_j-b_i)]} \quad \text{（公式2）}$$

其中，X_{ij}為受試者j在試題 i 的作答反應，答對記為1，答錯記為0；a_i為試題 i 之試題鑑別度參數（item discrimination parameter）；b_i為試題 i 之試題難度參數。

(3)三參數對數模式

在試題反應理論的3PL模式下，假定測驗會發生猜題之現象，故假設受試者 j 之能力為 θ_j，其作答試題 i 通過的機率如下(Birnbaum, 1968; Lord, 1980)：

$$P(X_{ij}=1|\theta_j,b_i,a_i,c_i)=c_i+\frac{(1-c_i)}{1+\exp[-a_i(\theta_j-b_i)]} \quad \text{（公式3）}$$

其中，X_{ij} 為受試者 j 在試題 i 的作答反應，答對記為1，答錯記為0；a_i 為試題 i 之試題鑑別度參數；b_i 為試題 i 之試題難度參數；c_i為試題 i 之試題猜測度參數(item guessing parameter)。

2.多點計分模式

(1)部分給分模式

部分給分模式(partial credit model, PCM)是由Masters (1982) 所提出，為Rasch’s model在多點計分的一個應用。

$$P_{ik}(\theta_j)=\frac{\exp\left[\sum_{v=1}^{k}(\theta_j-b_{iv})\right]}{\sum_{c=1}^{m_i}\left[\exp\sum_{v=1}^{c}(\theta-b_{iv})\right]} \quad \text{（公式4）}$$

其中，$P_{i1}=0$且$\sum_{c=1}^{1}(\theta_j-b_{iv})\equiv 0$；$\theta_j$表示受試者$j$的能力；$k$為受試者的回答所屬類別，$k=1\cdots m_i$；$m_i$為隨題目而變的變數；$m_i$是第 i 題所有的類別數；$P_{ik}(\theta_j)$ 表示能力為 θ_j 的受試者 j 在第 i 題得 k 類的機率$(0<P_{ik}(\theta_j)<1)$；b_{iv}：指第 i 題第 v 個的試題步驟難度參數(item step parameter)或類別閾參數(category intersection parameter)，隨著類別界線(category boundary)而變，相鄰在兩類別間，就有一個b_{iv}參數$(-\infty<b_{iv}<\infty)$，即b_{ik}為$P_{i,k-1}(\theta_j)$和$P_{ik}(\theta_j)$

的交點。在部分給分模式的公式中，我們可以發現如果試題爲二元計分，則用部分給分模式模式來分析試題會與使用Rasch單參數模式來分析相同。

3.廣義部分給分模式

廣義部分給分模式(generalized partial credit model, GPCM)是部分給分模式的延伸，由Muraki(1992)所提出，爲各試題之間有不同的鑑別度參數，廣義部分給分模式定義如下：

$$P_{ik}(\theta_j)=\frac{\exp\left[\sum_{v=1}^{k}a_i(\theta_j-b_{iv})\right]}{\sum_{c=1}^{m_i}\left[\exp\sum_{v=1}^{c}a_i(\theta_j-b_{iv})\right]}=\frac{\exp\left[\sum_{v=1}^{k}a_i(\theta-b_i+d_v)\right]}{\sum_{c=1}^{m_i}\left[\exp\sum_{v=1}^{c}a_i(\theta-b_i+d_v)\right]}\quad\text{（公式5）}$$

其中$d_1\equiv 0$，爲了在進行參數估計時，使其有一個相對原點，$b_{iv}=b_i-d_v$；θ_j 表示受試者 j 的能力；k 爲受試者的回答所屬類別，k=1…m_i；m_i爲隨題目而變的變數，m_i是第 i 題所有的類別數；$P_{ik}(\theta_j)$表示能力爲 θ_j的受試者 j 在第 i 題得 k 類的機率($0<P_{ik}(\theta_j)<1$)；$b_{iv}=b_i-d_v$，b_{iv}爲第 i 題第v 個的試題步驟難度參數(item step parameter)或類別閾參數(category intersection parameter)，隨著類別界線(category boundary)而變，相鄰在兩類別間，就有一個b_{iv}參數($-\infty<b_{iv}<\infty$)，即b_{ik}爲$P_{i,k-1}(\theta_j)$和$P_{ik}(\theta_j)$的交點，同一試題內的試題步驟參數不需是有序的；b_i爲試題 i 位置參數(item location parameter)；d_v爲閾參數(threshold parameter); d_k爲同一試題內的第 k 類和其他類別的相對難度(Andrich, 1982)；a_i：試題 i 的斜率參數，同一試題在各類別選項有相同的斜率參數，但不同的試題有不同斜率。

（二）多向度試題反應理論模式

目前常見的多向度試題反應理論(multidimensional item response theory, MIRT)大多是單向度試題反應模式(unidimensional item response theory, UIRT)的衍生模式。以下將介紹幾種常見的多向度試題反應理論模式，分別爲多向度隨機係數多項logit模式(MRCMLM)、多向度二參數模式(multidimensional two parameters model, M2PL)、多向度三參數模式(multidimensional three parameters model, M3PL)。

1.多向度二參數模式(M2PL)

多向度二參數模式爲二參數logistic模式(two-parameter logistic model, 2PL)所衍生的模式(Mckinley & Reckase, 1983；Reckase & Mckinley, 1991)，其模式定義如下：

$$P_i(x_{ij}=1\,|\,\mathbf{a}_i,d_i,\boldsymbol{\theta}_j)=\frac{1}{1+\exp[-(\mathbf{a}_i'\boldsymbol{\theta}_j-d_i)]} \qquad \text{（公式6）}$$

其中X_{ij}爲受試者作答反應型態，1表示答對該試題，0表示答錯該試題；a_i爲試題鑑別度向量，d_i爲試題難度，θ_j爲受試者能力向量。多向度二參數模式與二參數IRT模式差別爲將原本的受試者能力值θ與試題鑑別度a擴展爲向量θ_j及a_i，透過向量來表示，以將多向度的能力同時包含在模式中。由於試題鑑別度向量a_i包含多個向度的鑑別度，如此多個向度的鑑別度無法完整表現出單一試題的鑑別度，因此Reckase & McKinley(1991)定義出兩個常用的多向度指標，一個是第i題的多向度鑑別度參數(multidimensional discrimination parameter, $MDISC_i$) $MDISC_i=(\sum_{k=1}^{m}a_{ik}{}^2)^{\frac{1}{2}}$，其中$m$爲能力向度數目；另一個是第$i$題的多向度難度參數(multidimensional difficulty parameter, $MDISC_i$) $MDISC_i=\frac{-d_i}{MDISC_i}$；另外，爲了能具體觀察試題的向度結構，以顯示個別向度鑑別度a_{ik}與多向度鑑別度參數$MDISC_i$之間的關係，Ackerman(1996)定義試題所要測量的能力方向與各能力向度間的夾角$\cos\alpha_{ik}=\frac{a_{ik}}{MDISC_i}$，$k=1,\cdots,m_i$。

2.多向度三參數模式(M3PL)

多向度三參數模式爲三參數logistic模式(three-parameter logistic model, 3PL)的改良，將三參數logistic模式中的能力參數與鑑別度參數改成向量的型式(Hattie, 1981；Sympson, 1978)，模式定義如下：

$$P_i(U_i=1\,|\,\mathbf{a}_i,b_i,c_i,\boldsymbol{\theta}_j)=c_i+\frac{1-c_i}{1+\exp[-\mathbf{a}_i'(\boldsymbol{\theta}_j-b\mathbf{1})]} \qquad \text{（公式7）}$$

其中，U_i 爲第 i 題反應型態；θ_j 爲受試者能力向量；c_i 爲試題的猜測參數；a_i 爲試題鑑別度向量；而爲了使試題的難度成爲向量用以與能力向量相減，故將難度參數 b 與向量 1 相乘。多向度三參數的模式有其他表示法，由 Reckase(1997) 提出的多向度三參數洛基模式 (multidimensional three-parameter logistic model, M-3PL) 如公式 8 所示，能力爲 $\vec{\theta}_i$ 的受試者，在二元計分試題 j 的答對機率爲：

$$P_{ij1} = P(x_{ij}=1 \mid \vec{\theta}_i, \vec{\beta}_j) = \beta_{3j} + \frac{1-\beta_{3j}}{1+e^{(-\vec{\beta}_j \Theta \vec{\theta}_i^{T} + \beta_{1j})}} \qquad \text{（公式 8）}$$

其中，x_{ij} 爲受試者 i 在試題 j 的作答反應，答對時 $x_{ij}=1$，答錯時 $x_{ij}=0$；$\vec{\beta}_{2j} = (\beta_{2j1}, \cdots, \beta_{2jD})$ 爲 D 個向度的試題鑑別度參數向量；β_{1j} 爲試題難度參數；β_{3j} 爲試題猜測參數；$\vec{\beta}_{2j} \Theta \vec{\theta}_{2j}^{T} = \sum_{l=1}^{D} \beta_{2jl}\theta_{il}$；第 題的試題參數爲 $\vec{\beta}_j = \vec{\beta}_{2j}, \beta_{1j}, \beta_{3j}$。

3. 多向度隨機係數多項 logit 模式（MRCMLM）

多向度隨機係數多項洛基模式是由 Adams、Wilson 與 Wang(1997) 等人所提出，MRCMLM 爲 Rasch 模式的衍生模式，是一個混合的 co-efficients 模型（mixed co-efficients model），試題參數是由未知的參數所描述，而受試者的潛在變數 θ，是一個隨機變項，模式定義如下：

$$P(\mathbf{X}_{ik}=1;\mathbf{A},\mathbf{B},\boldsymbol{\xi}|\boldsymbol{\theta}) = \frac{\exp(\mathbf{b}'_{ik}\boldsymbol{\theta}+\mathbf{a}'_{ik}\boldsymbol{\xi})}{\sum_{k=1}^{K_i}\exp(\mathbf{b}'_{ik}\boldsymbol{\theta}+\mathbf{a}'_{ik}\boldsymbol{\xi})} \qquad \text{（公式 9）}$$

其中 $\mathbf{X}_{ik} = (\mathrm{X}_{i1}, \mathrm{X}_{i2}, ..., \mathrm{X}_{iK_i})'$，$k = 0,1,..., K_i$ 表示受試者反應型態，

$$\mathrm{X}_{iK} = \begin{cases} 1 & \text{爲 } i \text{ 第題作答第 } k \text{ 個反應類別} \\ 0 & \text{表其他} \end{cases};$$

$\boldsymbol{\xi}' = (\xi_1, \xi_2, ..., \xi_p)$ 爲試題參數向量（p 個參數）；θ=(θ_1, θ_2, ⋯, θ_D) 爲受試者的能力向量（D 個向度）；$' = (\mathbf{a}_{11}, \mathbf{a}_{12}, ...\mathbf{a}_{1K_1}, \mathbf{a}_{21}, \mathbf{a}_{22}, ...\mathbf{a}_{2K_2}, ..., \mathbf{a}_{nK_n})$ 爲整份測驗的設計矩陣；$\mathbf{a}_{ik}(i = 1,..., n \text{ and } k = 1,..., K_i)$ 爲第 i 題中第 k 個反應類別的設計向量，每個向量長度爲 p；$\mathbf{B} = (\mathbf{B}'_1, \mathbf{B}'_2, ..., \mathbf{B}'_n)'$：整份測驗的計分矩陣；$\mathbf{B}_i = (\mathbf{b}_{i1}, \mathbf{b}_{i2}, ..., \mathbf{b}_{iD})'$：第 i 題的計分子矩陣；$\mathbf{b}_{ik} = (b_{ik1}, b_{ik2}, ..., b_{ikD})'$：在 D 個向度中，第 i 題回答第 k 個反應類別的計分向量。

依據測驗架構又可分爲題間多向度測驗（between-item multidimensional test）與題內多向度測驗（within-item multidimensional test）兩種（Adams, Wilson & Wang, 1997），前者定義每個試題只測量一種能力，如圖1題間多向度；後者定義每個題目不只測量一種能力，如圖2題內多向度。

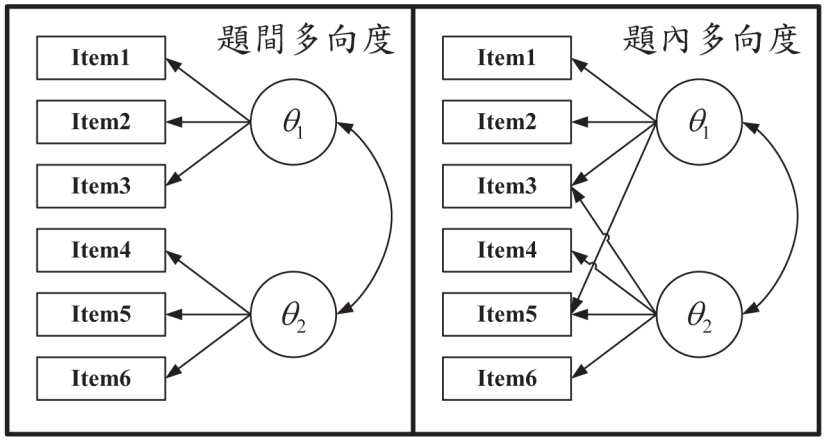

圖1 題間多向度評量架構圖　圖2 題內多向度評量架構圖

（三）高階層試題反應理論模式

隨著評量架構的日趨複雜，近年來更複雜的測驗理論相繼被提出。以PISA（The Programme for International Student Assessment）數學科評量架構爲例，除了數學科能力（mathematics），同時也希望得到學生在數量（quantity）、空間與形體（space and shape）、改變與關係（change and relationships）及不確定性（uncertainty）四個數學能力（OECD, 2005）。以往估計此種評量架構是以單向度試題反應理論估計數學科能力（如圖3），或以多向度試題反應理論估計四個數學能力（如圖4）。這樣的估計方式會產生較大的估計誤差，使用UIRT模式會違背其假設而使主要量尺能力（即數學科能力）估計不準確，或當次級量尺（即上述之數量、空間與形體、改變與關係及不確定性（uncertainty）四個數學能力）所對應的題數較少時，導致估計效果不好依據。有鑑於此，學者開始發展更複雜的測驗理論，Song(2007)提出一因子高階層IRT模式，同時包含整體能力(overall ability)與領域能力(domain ability)（如圖5），透過適當地參數估計過程可以同時獲得主要量尺能力和次級量尺能力的估計（de la Torre & Song, 2009）。

Song(2007)模擬研究顯示，當次級量尺之間不相依時，HO-IRT模式對主要量尺的估計結果會相似於UIRT模式；當彼此相依時，HO-IRT模式估計次級量尺會比UIRT模式更準確。

在HO-IRT模式中，一份測驗可觀察多個單向度子測驗(subtest)，也就是次級量尺$\theta_i^{(d)}$，$\theta_i^{(d)}$，表示第 i 位受試者在次級量尺 d 的表現，其中$d=1,2,3,\ldots,D$。當不同次級量尺均測量相同能力時，整份測驗即爲單向度測驗；若不同次級量尺間有關聯，會藉由高階層能力 θ_i來連結這些次級量尺，θ_i表示第 i 位受試者在主要量尺的表現，而次級量尺是主要量尺的線性函數，公式定義如下：

$$\theta_i^{(d)} = \lambda^{(d)}\theta_i + \varepsilon_{id}$$

其中$\lambda^{(d)}$爲迴歸參數且$\left|\lambda^{(d)}\right| \le 1$，$\varepsilon_{id}$爲誤差項服從平均數爲0、變異數爲$1-\lambda^{(d)2}$的常態分布，$\theta_i^{(d)}$服從平均數爲$\lambda^{(d)} * \theta_i^{(d)}$、變異數爲$1-\lambda^{(d)2}$的常態分布。$\lambda^{(d)}$爲主要量尺與次級量尺間的相關，$\lambda^{(d)} \times \lambda^{(d)}$表示次級量尺與次級量尺間的相關；$\lambda^{(d)}$可爲負數，但在教育測驗上，主要量尺與次級量尺皆爲正相關，故只考慮 $0 \le \lambda^{(d)} \le 1$ 。

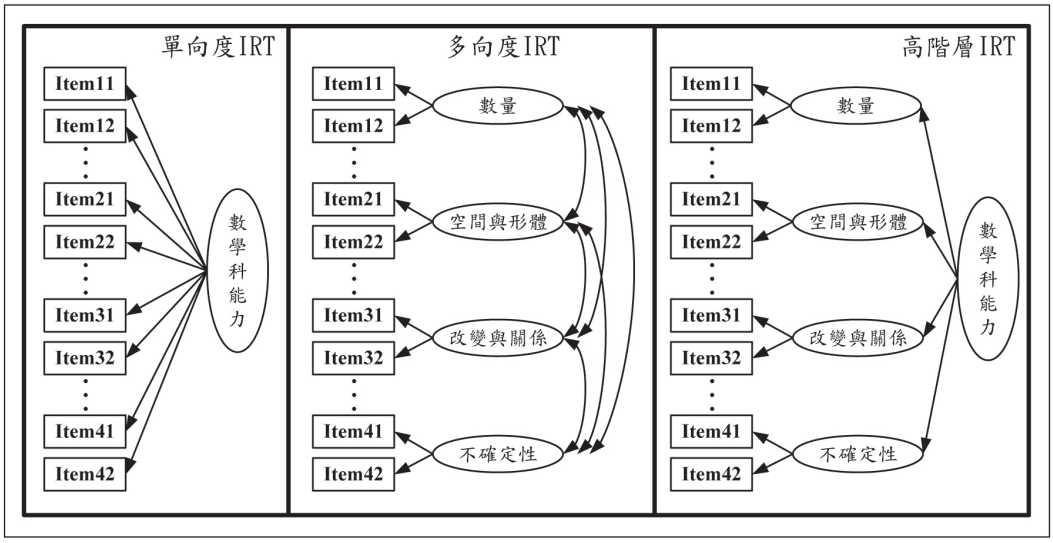

圖3　單向度IRT評量架構圖

圖4多向度IRT評量架構圖

圖5　一因子高階層IRT評量架構圖

三、試題反應理論之應用

（一）試題反應理論電腦程式簡介

1.電腦程式簡介

應用試題反應理論分析測驗資料時，必須估計所選用試題反應函數的參數，參數估計常涉及艱深難懂的數學公式及繁瑣的計算過程，若沒有電腦套裝程式的即時配合，則在應用上會受到限制；目前電腦科技突飛猛進，各種適用於試題反應理論的電腦軟體程式相繼問世，只要使用者學會這些程式，便能有效率的獲取所需要的參數估計值，進一步對測驗資料進行分析與解釋。目前常見的試題反應理論電腦程式整理如表1。表1的資料顯示，目前無論是在二元計分、多點計分、單向度IRT模式或多向度IRT模式，都已經開發相對應的電腦程式，但對於高階層試題反應理論模式只有部分學者於相關的學術論文中自行開發估計程式，仍無商業軟體或免費的電腦程式供讀者使用。

表 1
試題反應理論參數估計套裝軟體

軟體	可分析模式
BILOG-MG 3.0（Zimowski, Muraki, Mislevy, & Bock, 1996）	單、二、三參數對數模式（1PL、2PL、3PL）
MULTILOG7（Thissen, Chen & Bock, 1991）	單、二、三參數對數模式（1PL、2PL、3PL） Samejima's model for graded responses Bock's model for nominal (non-ordered) responses Steinberg's model for multiple-choice items
PARSCALE4.1（Muraki & Bock, 1997）	單、二、三參數對數模式（1PL、2PL、3PL） Samejima's model for graded responses Master's partial credit model Generalized partial credit model
NOHARM（Fraser, 1988）	潛在特徵模式（the latent trait models） 多向度二參數模式（M2PL） （Mckinley & Reckase, 1983; Reckase & Mckinley,1991） 多向度三參數模式（M3PL） （Hattie, 1981; Sympson, 1978） 維度多參數試題反應模式（Mensional multiparameter item）response（MMIR） （Beguin & Glas, 2001; Bock, Gibbons, & Muraki, 1988; McDonald, 1982; McKinley & Reckase, 1983; Muraki & Carlson, 1995） 多向度隨機係數多項洛基模式（multidimensional random coefficients multinomial logit model, MRCMLM） （Adams, Wilson, & Wang, 1997）
TESTFACT（Wilson, Gibbons, Schilling, Muraki & Bock, 2003）	多向度二參數模式M2PL （Mckinley & Reckase, 1983; Reckase & Mckinley,1991） 多向度三參數模式，M3PL （Hattie, 1981; Sympson, 1978） 維度多參數試題反應模式（mensional multiparameter item response,MMIR）

軟體	可分析模式
	（Beguin & Glas, 2001; Bock, Gibbons, & Muraki, 1988; McDonald, 1982; McKinley & Reckase, 1983; Muraki & Carlson, 1995） 多向度隨機係數多項洛基模式（multidimensional random coefficients multinomial logit model, MRCMLM） （Adams, Wilson, & Wang, 1997）
MAXLOG （Mckinley & Reckase, 1983）	多向度二參數模式（M2PL） （Mckinley & Reckase, 1983; Reckase & Mckinley,1991）
ConQuest （Wu, Adams, & Wilson, 1998）	多向度隨機係數多項洛基模式（multidimensional random coefficients multinomial logit model, MRCMLM） (Adams, Wilson, & Wang, 1997)
MIRTE 2 （Carlson, 1987）	多向度二參數模式（M2PL） （Mckinley & Reckase, 1983; Reckase & Mckinley,1991）
BMIRT （Yao, 2003）	多向度三參數模式（M3PL） 可分析混合二元計分試題與多點計分試題的測驗

（二）試題反應理論在大型教育測驗之應用

近年國內積極參加一些國際評比之大型測驗(large-scale assessments)，如PISA、國際數學與科學教育成就趨勢調查(Trends in International Mathematics and Science Study, TIMSS)，國際評比的成績收到國人的重視，而國家教育進展評量(National Assessment of Educational Progress, NAEP)則是較早也是較著名之大型測驗，這些大型測驗都是以試題反應理論爲主要測量模式，是應用試題反應理論最佳範例，以下將介紹這幾個測驗大型測驗並說明所使用的測量模式與等化設計。

1.國家教育進展評量(NAEP)

NAEP爲美國教育測驗服務社（Educational Testing Service, ETS）所發展的聯邦補助計畫，主要目的爲建立學生學習成就的趨勢。NAEP自1969年便開始定期地對4年級、8年級及12年級學生進行閱讀（reading）、數學（mathematics）、科學（science）、寫作（writing）之能力評量（NCES, 2005），是美國評量學生成就之代表，NAEP評量之範圍可分爲全國性的（National NAEP）、各州的（State NAEP）、

地區性的（NAEP Trial Urban District Assessment）評量（The Nation's Report Card, 2005；張鈿富、王世英、吳慧子、周文菁，2006）。NAEP之評量分爲主要評量（Main NAEP）與長期發展趨勢評量（Long-term Trend NAEP）兩類，主要的目的爲（1）反映學生在主要課程領域上應該知道和可以做的廣泛能力；（2）測量長時間範圍內的教育發展情形（張鈿富、王世英、吳慧子、周文菁，2006）。

2. 國際學生評量(PISA)

PISA測驗是由「經濟合作與發展組織（Organization for Economic Co-operation and Development, OECD）」主辦，目的在於了解個人參與社會活動的能力。主要的對象是15歲的學生，並進行其閱讀素養（reading literacy）、數學素養（mathematical literacy）、科學素養（scientific literacy）、及問題解決（problem solving）之能力評量。PISA每次進行評量會從數學、科學及閱讀三個領域中選定一個主要領域，例如：PISA 2000的主要領域爲閱讀，2003爲數學，2006爲科學，2009再回到閱讀，國內從第三次跨國學生評量測驗（PISA 2006）開始參與，一直延續至今。

3. 國際數學與科學教育成就趨勢調查(TIMSS)

TIMSS主要目的爲進行學生數學與科學教育成就趨勢調查研究，測試對象爲4年級與8年級之學生，TIMSS施測數學與科學兩學科，各學科分爲內容領域（content domain）與認知領域（cognitive domain）。欲評估學生能否掌握參與社會所需的知識與技能，並藉由國際評比來比較參與地區或國家的教育成效。自1999年進行TIMSS-R評量後，IEA(The International Association for the Evaluation of Education Achievement, IEA)計畫每隔四年辦理國際數學與科學教育成就研究一次，並改名爲TIMSS。

目前NAEP、TIMSS仍以UIRT爲主要使用測量模式，僅能對各個學科能力以單一能力值進行描述(Lee, Grigg, & Dion, 2007; Mullis, et al., 2007)，對各學科所屬之次級量尺(subscales)表現較無法作精確描述；PISA使用多向度試題反應理論(multidimensional item response theory, MIRT)中之多向度隨機係數多項logit模式(multidimensional random coefficients multinomial logit model, MRCML)進行測驗分析並對各學科之次級量尺進行估計；然而PISA使用多點計分模式對題組試

題進行分析(OECD, 2005)，未考慮題組試題對於參數估計之影響。Wang 和 Wilson(2005)研究結果顯示，如果測驗爲題組試題之測驗題型，忽略試題之間彼此可能相依之情形，則會高估能力參數且造成試題參數估計之偏差。

國內外的大型測驗，因題庫涵蓋不同認知程度及不同難度之試題，試題數量無法由單一受試者於短時間內完成，故多採用不同的等化設計進行，常見的等化設計是平衡不完全區塊設計（balanced incomplete block design, BIB）BIB設計是由Yates（1936）提出，並於1992年Rust & Johnson應用於測驗領域的題庫設計。此設計是指題庫中所有的試題區塊出現次數是相同的，且成對試題區塊出現於題本中的次數也必須是相同的。所謂的「平衡」是由於成對試題區塊出現於題本中的次數是相同的，因此在成對試題區塊平均數間之比較有相同的精準度。各題本中的試題區塊可能部分相同或完全不同，但是每一個試題區塊在所有題本中出現的次數是一樣的，亦即題庫中的每個試題所受測的學生約爲相同的。（Kuehl, 2000；郭伯臣、曾建銘、吳慧珉，2012）。以下介紹NAEP、TIMSS與PISA之等化設計。

1.NAEP

以1998年4年級公民爲例，使用之題本設計爲BIB設計，設計中共包含了6個試題區塊（M1~M6）組合成18個題本（S1~S18），爲了使試題區塊在題本前後出現的次數一致，故將題本16到18與題本13到15的兩個試題區塊作交換後組成（Andrew & Terry, 2001），以表2作說明。

表2

NAEP 1998年4年級公民題本區塊設計表

題本	區塊I	區塊II	題本	區塊I	區塊II
S1	M1	M2	S10	M4	M6
S2	M2	M3	S11	M5	M1
S3	M3	M4	S12	M6	M2
S4	M4	M5	S13	M1	M4
S5	M5	M6	S14	M2	M5
S6	M6	M1	S15	M3	M6
S7	M1	M3	S16	M4	M1
S8	M2	M4	S17	M5	M2
S9	M3	M5	S18	M6	M3

2.TIMSS

以2007年之題本設計爲例，每個題本由四個試題區塊組合而成，包含數學（M01~M14）與科學（Q01~Q14）各兩個試題區塊，爲了連結不同題本，每個試題區塊在題本中出現2次（Graham, Christine, Alka, & Ebru, 2008）。表3爲TIMSS2007年之題本區塊設計。

表3

TIMSS2007年題本區塊設計表

題本	區塊（Part I）		區塊（Part II）	
S1	M01	M02	Q01	Q02
S2	Q02	Q03	M02	M03
S3	M03	M04	Q03	Q04
S4	Q04	Q05	M04	M05
S5	M05	M06	Q05	Q06
S6	Q06	Q07	M06	M07
S7	M07	M08	Q07	Q08

題本	區塊（Part I）		區塊（Part II）	
S8	Q08	Q09	M08	M09
S9	M09	M10	Q09	Q10
S10	Q10	Q11	M10	M11
S11	M11	M12	Q11	Q12
S12	Q12	Q13	M12	M13
S13	M13	M14	Q13	Q14
S14	Q14	Q01	M14	M01

資料來源：TIMSS2007 Technical Report（p.34）

3.PISA

以PISA2006年之題本設計爲例，題本設計爲BIB設計，共包含13個題本（S1~S13），每個題本包含4個試題區塊（區塊I~區塊IV），每個試題區塊在題本中出現4次（r＝4），以及成對試題區塊在各題本中出現1次（λ＝1）（OECD, 2009），表4爲PISA2006年之題本區塊設計，其中試題區塊M1~M4代表數學科之試題區塊；Q1~Q7代表科學之試題區塊；R1~R2代表閱讀之試題區塊，每個題本內可能包含數學、科學或閱讀三種不同科目之試題區塊。

表4

PISA2006年題本區塊設計表

題本	區塊I	區塊II	區塊III	區塊IV
S1	Q1	Q2	Q4	Q7
S2	Q2	Q3	M3	R1
S3	Q3	Q4	M4	M1
S4	Q4	M3	Q5	M2
S5	Q5	Q6	Q7	Q3
S6	Q6	R2	R1	Q4
S7	Q7	R1	M2	M4

題本	區塊I	區塊II	區塊III	區塊IV
S8	M1	M2	Q2	Q6
S9	M2	Q1	Q3	R2
S10	M3	M4	Q6	Q1
S11	M4	Q5	R2	Q2
S12	R1	M1	Q1	Q5
S13	R2	Q7	M1	M3

資料來源：PISA2006 Technical Report（p.29）

在NAEP、TIMSS及PISA中，主要關注的焦點是母群或母群中某些群體之能力表現，這些大型測驗以可能值方法(plausible value methodology)進行群體統計描述，例如群體之平均數或標準差（Allen, Carlson, Johnson, & Mislevy, 1999; Foy, Galia, & Li, 2008; OECD, 2009），假如研究者想要瞭解不同群體的能力表現，則納入群體的背景變項進行可能值方法估計，藉以提升群體參數估計的精確度（Adams, Wilson & Wu, 1997）。國內許多大型測驗相關研究，多未使用可能值方法進行分析，而是直接計算個別受試者能力值的平均與變異，並將其視爲母群或個別群體的表現與其分散情形，再進一步的進行假設檢定，依據相關研究（Mislevey, 1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; OECD, 2009; Lee, et al., 2007）顯示：此種集合個體的能力值估計群體特性的方式將會產生嚴重的偏誤，故對於次級資料分析者，應使用這些大型測驗的使用手冊，正確使用這些大型測驗釋出的可能值進行相關的研究探討。

（三）以試題反應理論爲基礎之電腦化適性測驗

傳統紙筆測驗，無論受試者能力高低，都必須將一份試卷全部作答，往往發生高能力受試者感到試題太簡單而浪費時間，低能力受試者感到試題太困難而猜答，影響到測驗的準確性。電腦化適性測驗是逐題依據受試者作答反應，作爲選取下一道試題施測的依據，符合受試者能力來施測，使受試者感到試題難易適中，減少施測題數，節省施測時間。

由於網路迅速發展與電腦週邊設備大眾化，電腦的影音資料傳輸與呈現可以快速且穩定，因此在網路上實施測驗是具體可行的；透過電腦進行測驗擁有快速資料處理的優點，使得測驗的歷程、統計、分析與結果能自動詳細記錄，所以電腦化測驗的發展愈來愈受矚目且持續被研究中。在1960年時，美國陸軍總署、人事管理局、及其他聯邦機構，均大力支持贊助有關適性測驗的研究，舉辦特殊的專題研討會，並且有數百篇的相關研究論文發表且集結成冊（Wainer, 2000; Weiss,1983; 余民寧，1997）。在國外許多商業機構、測驗公司，也陸續有測驗產品發行，如Larson & Smith (1988) 所發展的「西班牙電腦化適性安置測驗」(A Spanish Computerized Adaptive Placement Exam)；Assessment Systems公司所發行的「MicroCAT適性測驗系統」與「Minnesota文書性向評鑑測驗」；美國心理測驗公司（The Psychological Corporation）所發行「中學區分性向測驗」(李茂能，2000)。目前電腦化適性測驗在大型測驗的GRE (Graduate Record Examinations)、TOFEL (Test of English as a Foreign Language)、GMAT (Graduate Management Admission Test) 等也均已實施。

電腦適性測驗是爲受試者「量身打造」的個別測驗，符合「因才施測」、「經濟有效」、「誤差最小」的原則（李茂能，2000），主要優點有：

1. 施測試題的難度，符合受測者的能力範圍內，因此試題具備適性。
2. 受試者的施測試題不同，在適性測驗的過程中，若符合測驗終止條件則結束測驗，受測者不用作答偏易或偏難的試題，節省施測時間。
3. 適性測驗中，能對估計精準度提供最大訊息量的試題會被優先選取爲施測試題，因此受試者能力估計的精準度最高。

在測驗編制過程以IRT爲基礎，能經由分析試題獲得篩選過試題參數的題庫，以確保題庫品質的好壞；在測驗施測過程以IRT爲基礎的CAT，能針對不同受試者提供適合受試者的試題，在可容忍的誤差下，讓測驗的時間大幅縮短，因此以IRT爲基礎的適性診斷系統也陸續在研發（王雯芳，2004；陳新豐，2004；黃吉楠，2004；楊踶齊，2006；蔡文龍，2008；蕭顯勝、黃啓彥、游光昭，2006）。

以IRT爲基礎的電腦化適性測驗，包含了五項基本要素：測驗題庫、測驗起點、能力估計、選題策略與測驗終止條件（Wainer, 2000），而依照CAT的不同類型又可分作單向度CAT與多向度CAT兩大類，本研究針對不同CAT實施過程所需的要素與CAT施測流程說明如圖6所示：

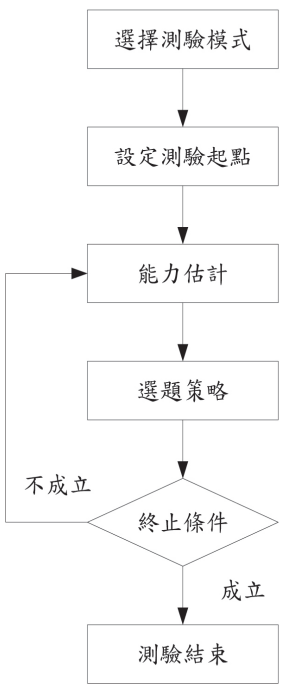

圖6 CAT施測流程

1. 測驗題庫

電腦化適性測驗題庫好壞通常以題庫大小與試題參數來評估，若CAT施測長度爲傳統紙筆測驗長度的一半，則CAT題庫大小最好是傳統紙筆測驗長度的6至8倍，也就是說題庫大小至少爲CAT施測長度的12倍(Stocking, 1994)，當題庫長度爲3倍以上，精確度與作答效率才有顯著差異(Hung, 1988)。對單向度CAT的3PLM來說，一個好的題庫其試題鑑別度應大於0.8，試題難度應該與受試者的母群能力分布相近，試題猜測度應小於0.25（王寶塘，1995）。Ree(1981)曾以最大訊息法爲選題策略的研究，在沒有曝光率控管下，題庫長度大於200題時，對能力估計的精準度並不會明顯增加。對多向度CAT來說，Wang, Chen 與Cheng (2004) 的研究顯示，當題庫向度之間爲高相關時，多向度IRT分析可以大幅提高各向度的信度，由原本的0.6（單向度IRT分析）提昇至0.8。

2. 測驗起點

CAT是依照受試者能力挑選適當試題對受試者進行施測，在測驗初始階段中，受試者能力高低未知下，必須決定測驗起始點，以挑選第一道試題提供受試者進行施測。常用的決定起始試題的方法（王寶塘，1995；陳麗如，1998；錢永財，2006；Chang & Ansley, 2003）分別介紹如下：

(1)中等難度題目：先假設受試者爲中等能力，在題庫中挑選適合中等能力難度的試題作爲施測的起始試題，若每位受試者都使用相同的試題，則起始試題的保密性需要特別考量。

(2)依據受試者能力選題：由受試者的基本資料（年齡、學習經驗或其他測驗結果）估算受試者的能力初始值，再利用能力初始值決定測驗的起始試題。

(3)自由選題：由受試者在接受測驗的時候，自行判定自己的程度，以決定施測的起始題，但容易受到受試者主觀判斷的影響。

(4)隨機選題：由電腦隨機選題，但一般限定試題選取範圍不可超過題庫本身。Lord (1977) 發現不同起始點對於測驗標準誤(standard error of measurement)並沒有很大差別。

(5)隨機法：McBride與Martin(1983)發表隨機法。隨機法是在施測前期使用隨機選取策略來避免題目的過度曝光，實施方法爲對每一個受試者，依能力初始值從題庫中選出5個訊息量最大的題目，

從這5題中隨機選取出一題施測並重新估計能力值；依新估計的能力值從題庫中選出4個訊息量最大的題目，再從這4題中隨機選取出一題施測並重新估計能力值；重複相同模式，選取施測的前4題，第5題之後就使用最大訊息法來選取施測題目。此外根據Chang與Ansley (2003) 的研究指出隨機法在不同題庫中的比較，試題最大曝光率皆介於0.64~0.74，遠高於可接受試題最大曝光率。

(6)初始階段b值分層隨機選題法：錢永財（2006）提出初始階段b值分層隨機選題法來進行測驗初期選題，其實施步驟爲步驟一：將題庫依b值大小分成k層，k爲初始階段選題的題數；步驟二：施測前k題時，分別自k層中各自隨機選取一個試題，進行施測，研究顯示，此方法可有效降低試題最大曝光率、有效提高題庫的使用率。

3. 能力估計

3.1單向度CAT

CAT常見的能力估計方法有最大概似估計法、期望後驗估計法與最大後驗估計法，三種能力估計方法分別介紹如下：

(1)最大概似估計法

最大概似估計法（maximum likelihood estimation, MLE）是設測驗共有 個試題，試題間彼此獨立，概似函數定義如下：

$$L(\mathrm{u}\mid\theta)=L(X_1,...,X_n\mid\theta)=\prod_{i=1}^{n}P_i^{X_i}Q_i^{1-X_i}\ ，\qquad（公式10）$$

其中，u爲所有作答反應的向量，$L(X_1,...,X_n\mid\theta)$爲概似函數（likelihood function）；θ爲受試者的真實能力；X_i指受試者在第i題的作答反應，答對爲1，答錯爲0；P_i指受試者在第i題的答對機率；Q_i指受試者在第i題的答錯機率。爲了加速找到概似函數的最大值，通常是先對概似函數取對數，再以牛頓-約佛森（Newton-Raphson）法來進行迭代。使用MLE能力估計的公式如公式11所示，而第j次的能力估計的變動量爲$\delta^{(j)}$如公式12所示：

$$\theta^{(j)}=\theta^{(j\text{-}1)}-\delta^{(j)}\ ，\qquad（公式11）$$

$$\delta^{(j)}=[\frac{\partial^2 \ln L(u|\theta)}{\partial\theta^2}]^{-1}\times\frac{\partial \ln L(u|\theta)}{\partial\theta}\ ,\qquad\text{（公式12）}$$

(2)期望後驗估計法

Bock和Mislevy (1982) 提出期望後驗法（expected a posteriori，EAP）是尋找能力值的事後機率密度函數的期望值，公式定義如下：

$$\theta_{EAP}=\sum_{q=1}^{k_q}\theta_q f(\theta_q|\mathbf{U})=\sum_{q=1}^{k_q}\theta_q\frac{L(\mathbf{U}|\theta_q)f(\theta_q)}{\sum_{q=1}^{k_q}[L(\mathbf{U}|\theta_q)f(\theta_q)]}\ ,\qquad\text{（公式13）}$$

其中U爲所有作答反應的向量，$L(\mathbf{U}|\theta_q)$爲概似函數（likelihood function）；θ_q 爲受試者的真實能力；q是計算能力的期望值時所切割成的分割點（quadrature point），共有k_q 點，k_q愈大，計算的愈精確。不過這種估計方法不需要使用牛頓-約佛森法來進行迭代，而且隨著所選取的分割點數愈多，所需的計算量較龐大，計算時間也比較久。

(3)最大後驗估計法

貝氏最大後驗法(maximum a posteriori, MAP)是以受試者的事前能力分布$f(\theta)$作爲加權值，形成事後機率密度函數，並找出能使此事後機率密度函數最大化的程度值，稱爲MAP。事後機率密度函數定義如下：

$$f(\theta|\mathbf{U})=\frac{L(\mathbf{U}|\theta)f(\theta)}{f(\mathbf{U})}\ ,\qquad\text{（公式14）}$$

其中，$L(\mathbf{U}|\theta)$是受試者 θ 的概似函數， $f(\mathbf{U})$是受試者的邊際機率，是由$L(\mathbf{U}|\theta)f(\theta)$從$-\infty\sim\infty$積分所得，爲了加速找到事後機率密度函數的最大值，通常也是以牛頓-約佛森（Newton-Raphson）法來進行迭代。

(4)各種能力估計方法的比較

雖然MLE有不錯的估計效能，但有實務上的限制，當受試者作答反應爲全對或全錯時，MLE無法估計受試者能力值(Wang & Vispoel, 1998)。而EAP或MAP可以估計全對或全錯的作答反應，但若事前分配不正確，則能力估計偏差將會很大(Baker & Kim, 2004)。洪碧霞、吳裕益、吳鐵雄、陳英豪（1992）作過各種能力估計方法的比較，MLE比較沒有迴歸性的偏誤（bias），但均方根誤（root mean square of error, RMSE）較大；EAP與MAP有迴歸性的偏誤，但均方根誤較小。

3.2多向度CAT

多向度CAT常見的能力估計方法有最大概似估計法、期望後驗估計法與最大後驗估計法，與單向度CAT的估計法類似，主要由估計單一個能力值變成同時估計多個能力值，三種能力估計方法介紹如下：

(1)最大概似估計法

Segall (1996) 提出多向度的MLE是設測驗共有n個多向度的試題，試題間彼此獨立，概似函數定義如下：

$$L(\mathbf{u}\,|\,\theta)=L(u_1,u_2,\ldots,u_n\,|\,\theta)=\prod_{i\in v}P_i(\theta)^{u_i}Q_i(\theta)^{1-u_i}\ ,\qquad\text{（公式15）}$$

其中$L(\mathbf{u}\,|\,\theta)$爲概似函數（likelihood function）；θ爲受試者的真實能力向量，θ =[θ_1，θ_2，…，θ_D；u_i指受試者在第i題的作答反應，答對爲1，答錯爲0；P_i指受試者在第i題的答對機率；Q_i指受試者在第i題的答錯機率。

爲了加速找到概似函數的最大值，通常是先對概似函數取對數，再以牛頓-約佛森（Newton-Raphson）法來進行迭代，MLE能力估計的公式如公式16所示，而每次能力估計的變動量爲$\delta^{(j)}$如公式17所示（Wang, 1994）：

$$\theta^{(j)}=\theta^{(j\text{-}1)}-\delta^{(j)}\ ,\qquad\text{（公式16）}$$

$$\delta^{(j)}=\left[\frac{\partial^2 \ln L(\mathrm{u}\,|\,\theta)}{\partial\theta\partial\theta}\right]^{-1}\times\frac{\partial \ln L(\mathrm{u}\,|\,\theta)}{\partial\theta}\ ,\qquad\text{（公式17）}$$

$$\text{其中}\ \frac{\partial}{\partial\theta}\ln L(\theta|\mathbf{u})=\begin{bmatrix}\frac{\partial}{\partial\theta_1}\ln L(\mathbf{u}|\theta)\\ \frac{\partial}{\partial\theta_2}\ln L(\mathbf{u}|\theta)\\ \vdots\\ \frac{\partial}{\partial\theta_D}\ln L(\mathbf{u}|\theta)\end{bmatrix}，\qquad（公式18）$$

$$\text{所以}\ \frac{\partial\ln f(\boldsymbol{\theta}|\mathbf{u})}{\partial\theta_r}=\frac{\partial}{\partial\theta_r}\ln L(\mathbf{u}|\boldsymbol{\theta})=\sum_{i\in n}(u_i-P_i(\theta))，\qquad（公式19）$$

若第 i 題試題只測量第 r 種能力向度，則 $\frac{\partial[P_i(\theta)^{u_i}Q_i(\theta)^{(1-u_i)}]}{\partial\theta_r}=u_i-P_i(\theta)$

，若第 i 題試題沒有測量第 r 能力向度，則 $\frac{\partial[P_i(\theta)^{u_i}Q_i(\theta)^{(1-u_i)}]}{\partial\theta_r}=0$ ，

$$\frac{\partial^2}{\partial\theta_r\partial\theta_s}\ln L(\mathbf{u}|\boldsymbol{\theta})=-\sum_{i\in n}(P_i(\theta)\times Q_i(\theta))，\qquad（公式20）$$

$$\mathrm{J}(\theta)=\frac{\partial^2\ln f(\theta|\mathbf{u})}{\partial\theta\partial\theta}=\begin{bmatrix}\frac{\partial^2\ln f(\theta|\mathbf{u})}{\partial\theta_1^2} & \cdots & \cdots & \frac{\partial^2\ln f(\theta|\mathbf{u})}{\partial\theta_1\partial\theta_D}\\ & \frac{\partial^2\ln f(\theta|\mathbf{u})}{\partial\theta_2^2} & \cdots & \frac{\partial^2\ln f(\theta|\mathbf{u})}{\partial\theta_2\partial\theta_D}\\ & & & \vdots\\ & & & \frac{\partial^2\ln f(\theta|\mathbf{u})}{\partial\theta_{\mathrm{D}}^2}\end{bmatrix}，（公式21）$$

若第 i 題試題只測量第 r 種能力向度，則 ，$\frac{\partial[P_i(\theta)^{u_i}Q_i(\theta)^{(1-u_i)}]}{\partial\theta_r\partial\theta_r}=-[P_i(\theta)\times Q_i(\theta)]$，

若第 i 題試題沒有測量第 r 種能力向度，則 $\frac{\partial[P_i(\theta)^{u_i}Q_i(\theta)^{(1-u_i)}]}{\partial\theta_r\partial\theta_r}=0$ 。

(2)期望後驗估計法

多向度的EAP是將單向度的EAP能力值考慮能力向量，公式定義如下：

$$\theta_{EAP}=\sum_{q=1}^{k_q}\theta_q f(\theta_q\mid\mathbf{U})=\sum_{q=1}^{k_q}\theta_q\frac{L(\mathbf{U}\mid\theta_q)f(\theta_q)}{\sum_{q=1}^{k_q}[L(\mathbf{U}\mid\theta_q)f(\theta_q)]}\text{，（公式22）}$$

其中 q 是計算能力的期望值時所切割成的分割點（quadrature point），q 愈多使得計算愈精確；$f(\theta_q)$ 是多變量常態分配，公式定義如下：

$$f(\theta_1,\cdots,\theta_p)=\frac{1}{(2\pi)^{p/2}\left|\Phi\right|^{1/2}}\exp[-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})'\Phi^{-1}(\mathbf{x}-\boldsymbol{\mu})],-\infty<x<\infty\text{，（公式23）}$$

其中 $\theta'=[\theta_1,\cdots,\theta_D]$，$\mu$，為 θ 的平均數向量，$\boldsymbol{\mu}'=[\mu_1,\cdots,\mu_D]$，$\Phi$ 為 θ 的共變數矩陣，若變數矩陣標準化後，即為相關矩陣，

$\Phi=\begin{bmatrix}\sigma_1^2\cdots 0\\ \cdots\cdots\\ 0\cdots\sigma_D^2\end{bmatrix}$，在多向度中，分割點的數量會成向度數的指數倍增加，當向度數增加，則能力估計時間就會拉長，若減少各向度的分割點，又會降低能力估計的精準度。

(3)最大後驗估計法

Segall(1996)提出多向度的MAP，它的事後機率密度函數與EAP相同。為了加速找到事後機率密度函數的最大值，MAP比照MLE依牛頓-約佛森程序來進行。首先將分別對d個能力向度進行偏微分。MRCMLM事後機率度函數的一階偏微分向量中的元素如公式24所示，與二階偏微分向量中的元素如公式25（陳柏熹，2000；Wang, 1994）所示：

$$\frac{\partial\ln f(\theta|\mathbf{u})}{\partial\theta_d}=\frac{\partial}{\partial\theta_d}\ln L(\mathbf{u}|\theta)-\frac{\partial}{\partial\theta_d}[(\theta-\mathbf{u})'\Phi^{-1}(\theta-\mathbf{u})]$$

$$=\sum_{i\in V}(u_i-P_i(\theta))-\frac{\partial}{\partial\theta_d}[(\theta-\mathbf{u})'\Phi^{-1}(\theta-\mathbf{u})]\text{，}\qquad\text{（公式24）}$$

其中u為θ的平均數向量，Φ為θ的共變數矩陣，

$$\frac{\partial^2}{\partial\theta_r\partial\theta_s}\ln L(\mathbf{u}|\boldsymbol{\theta})=-\sum_{i\in n}(P_i(\theta)\times Q_i(\theta))-\phi^{rs}\ ,\qquad\text{（公式25）}$$

其他程序則比照最大概似估計法來進行。

(4)各種能力估計方法的比較

陳柏熹（2006）研究指出MLE、EAP、MAP這三種方法在多向度CAT中各有其優缺點，雖然從整體信度與測量誤差而言，MAP是比較好的，但是迴歸性的偏誤也是最嚴重的；EAP估計精準度與MAP差不多，但當向度數較高時，能力估計所需的時間就會太久；MLE是估計精準度較差的方法。所以建議當能力向度數少於四個向度可以使用EAP，以減少迴歸性偏誤的問題；但是當能力向度達到四個或四個以上時，最好使用MAP來進行，不過需要注意迴歸性偏誤的問題。

4.選題策略

4.1單向度CAT

選題法是電腦適性測驗中重要的要素之一，不同選題法會導致不同的測驗效率，常用的選題法介紹如下：

(1)最大訊息法

本研究中，試題的選取方法最常使用為最大訊息法，其實施步驟有步驟一：假設受試者目前能力估計值為$\hat{\theta}$，依據$\hat{\theta}$計算尚未施測試題的訊息量，計算式子參考公式9-4所示；步驟二：選取試題訊息量最大的試題，當作下一施測題目。

(2)最接近偏移難度法

若猜測度$c_j\neq 0$時，試題訊息最大值不會發生在難度b_j，會產生偏移至m_j，最接近偏移難度法為選擇題目偏移難度最接近受試者能力估計值$\hat{\theta}$的題目，作為下一階段施測的題目。偏移難度m_j(Birnbaum, 1968) 定義如下：

$$m_j=b_j+\frac{1}{d\cdot a_j}\log\left(\frac{1+\sqrt{1+8c_j}}{2}\right)\ ,\qquad\text{（公式26）}$$

則選題時選擇尚未施測且選題函數F_j最小的題目，選題函數定義如下：

$$F_j(\hat{\theta}) = |\hat{\theta} - m_j|，\qquad \text{（公式27）}$$

(3)區間式最大訊息法

區間式最大訊息法使用區間能力值的題目訊息量加總，來取代在某一點能力值的題目訊息量(Veerkamp & Berger, 1997)。區間式最大訊息法是選擇訊息函數在信賴區間內的面積，選擇最大的訊息面積，作爲下一題施測試題，所以選題時選取尚未施測且選題函數最大者，其函數定義如下：

$$F_j(\theta) = \int_{\hat{\theta}_l}^{\hat{\theta}_u} I_j(\theta)d\theta \qquad \text{（公式28）}$$

其中，$(\hat{\theta}_l, \hat{\theta}_u) = (\hat{\theta} - \frac{1.96}{\sqrt{I_T(\hat{\theta})}}, \hat{\theta} + \frac{1.96}{\sqrt{I_T(\hat{\theta})}})$；$I_j(\hat{\theta}) = \frac{[P_j'(\theta)]^2}{P_j(\theta)\cdot Q_j(\theta)}$。

(4)考慮 *b* 參數的 *a* 分層法

Chang, Qian & Ying(2001)提出考慮 *b* 參數的 *a* 分層法，希望將 *a* 分層法各分層中的 *b* 值分佈保持一致性，來打破 *a* 、*b* 之間的相關性。其實施步驟如下（假設*L* 爲測驗長度；*a* 爲試題鑑別度；*b* 爲試題難度）:

步驟一：將題庫依照 *b* 值由小至大分成*T*個區塊，第1個區塊包含最小的 *b* 值，第T個區塊包含最大的 *b* 值。

步驟二：在第t區塊（*t*=1,2,3,…*T*），將題庫依照 *a* 值由小至大分成 *K* 層，每一層包含一道試題，第1層包含最小的 *a* 值，第 *K* 層包含最大的 *a* 值。

步驟三：依序將每個區塊的第*K* 層（*k*=1,2,3,…*K*）合併成一層，因此題庫變成一具有*K* 層的結構。

步驟四：建立好試題的結構後，受試者依序從第1層開始施測，每一層選取*L*/*K* 個試題施測，一直施測到第*K* 層，直到受試者施測*L* 題。

謝友詩、劉湘川、郭伯臣（2006）研究指出在固定測驗長度下，能力均方根差由小而大排序分別爲考慮 *b* 參數的 *a* 分層法、最接近偏移難度法、鄰近法、區間式最大訊息法；考慮 *b* 參數的

a 分層法較區間式最大訊息法有較低的最大曝光率；在相同能力估計精準度下，考慮 b 參數的 a 分層法較區間式最大訊息法有較低的題目重複率。

(5) a-鄰近法

錢永財、劉家惠、郭伯臣（2005）提出改進鄰近法的 a -鄰近法，其第一步驟爲單點式最大訊息法，第二步驟改由控制 a 值，使早期能力值未準確時選用 a 值較低試題，a -鄰近法的實施步驟如下：

步驟一：將題庫依 b 值分三層，測驗前期三題採取隨機選題，使受試者在測驗前期施測難易度相差較大試題。

步驟二：估計初始化能力值估計值 $\hat{\theta}$ 。

步驟三：根據 $\hat{\theta}$ 選擇題庫中（測驗長度-已測驗題數）個訊息函數較大者的試題。

步驟四：再從（測驗長度-已測驗題數）個試題中，選其中題目 a 值最小測驗。

步驟五：重新估計能力值爲 $\hat{\theta}$ ，回到步驟三，直到測驗題數結束。

根據錢永財、劉家惠、郭伯臣（2005）的研究發現，使用 a-鄰近法，當題庫越大時，在試題曝光率的均勻度能越接近鄰近法，且能力估計誤差較低於鄰近法。

4.2 多向度CAT

Segall(1996)將概似函數取對數的二階偏微分透過公式29，以費雪訊息函數取代。

$$\mathrm{E}[\mathrm{H}(\theta^{(j)})] = -\mathrm{I}(\theta) \ , \qquad \text{（公式29）}$$

其中 $\mathrm{I}(\theta)$ 是費雪訊息矩陣，矩陣中第 r 列、第 s 行的元素表示如下：

$$I_{rs}(\theta_{(m)}) = -\mathrm{E}\left[\frac{\partial^2 \ln L}{\partial\theta_r \partial\theta_s}\right] = \sum_{i\in n} \frac{\dfrac{\partial P_i(\theta_{(m)})}{\partial\theta_r} \times \dfrac{\partial P_i(\theta_{(m)})}{\partial\theta_s}}{P_i(\theta_{(m)})Q_i(\theta_{(m)})} \ , \qquad \text{（公式30）}$$

上式表示受試者在施測完 m 個試題後，能力估計值爲 $\theta_{(m)}$ 的費雪訊息矩陣。而累加的第 i 題訊息量表示爲 $I_m(\theta_{(m)}, u_i)$ ，其定義如下：

$$I_{rs}(\theta_{(m)}, u_i) = \frac{\dfrac{\partial P_i(\theta_{(m)})}{\partial\theta_r} \times \dfrac{\partial P_i(\theta_{(m)})}{\partial\theta_s}}{P_i(\theta_{(m)})Q_i(\theta_{(m)})} \ , \qquad \text{（公式31）}$$

在多向度的CAT中，並非依據各單一向度的最大值來選題，而是讓費雪訊息矩陣的行列式值最大化的試題，其公式定義如下：

$$\left|\mathbf{I}(\boldsymbol{\theta}_{(m)})+\mathbf{I}(\boldsymbol{\theta}_{(m)},u_i)\right|，\qquad（公式32）$$

其中$\mathbf{I}(\theta_{(m)})$表示施測完前 m 題之後的訊息矩陣，$\mathbf{I}(\theta_{(m)},u_i)$是表示題庫內剩餘試題在能力估計向量$\theta_{(m)}$上的訊息矩陣。使用MAP時，其選題策略修正如下：

$$\left|\mathbf{I}(\boldsymbol{\theta}_{(m)})+\mathbf{I}(\boldsymbol{\theta}_{(m)},u_i)+\Phi^{-1}\right|，\qquad（公式33）$$

比較公式32與公式33，可以發現只差在能力先驗分配共變異數矩陣的反矩陣Φ^{-1}。

5.測驗終止條件

測驗終止條件主要分爲「最大測驗長度」與「能力估計的最小變動量」兩種，「最大測驗長度」是指測驗的題數達到預設的長度即停止測驗；「能力估計的最小變動量」是指當測驗的能力估計的變動量小於預設值即停止測驗。

四、結語

古典測驗理論雖然有許多限制，但由於理論模式簡單易懂，廣受一般社會大受的接受與喜愛，而試題反應理論架構嚴謹，但也相對艱澀難懂，本文首先比較古典測驗理論與試題反應理論之差異，並介紹目前常見的試題反應理論模式，包含單向度試題反應理論、多向度試題反應理論與較新的一因子高階層試題反應理論模式，並提供相對應的電腦程式供讀者參考使用。自1990年代之後，許多大型測驗也都應用試題反應理論進行量尺化程序，本文亦提供大型測驗應用IRT之例子供讀者參考，而隨著電腦科技的進步，電腦化適性測驗測驗已成爲實施測驗之新趨勢，本文針對電腦化適性測驗之理論與實施流程作一詳細介紹。試題反應理論之數學模式較複雜，讀者如能具備一些數學背景，閱讀本文較能得心應手。本文主要是介紹試題反應理論的基本概念與應用，許多試題反應理論所探討的課題與應用範疇是本文未提及的，如試題差異功能、參數估計、題組模式等，讀者可根據本文所提供的參考文獻或一些介紹試題反應理論的書籍，作延伸性之閱讀，將可更瞭解此一理論之內涵。試題反應理論至今仍是測驗領域之發展主軸，隨著新型測驗與更複雜評量架構之誕生，一些新的試題反應理論模式仍不斷被提出，有興趣的讀者可參考測驗領域的期刊，如Applied Psychological Measurement、Journal of Educational Measurement、Psychometrika等，將可獲得較新的試題反應理論資訊。

參考文獻

中文部份

王雯芳（2004）。網路測驗系統之建置--應用電腦化適性測驗於國民小學自然科技領域。華梵大學資訊管理學系碩士論文，未出版，台北縣。

王寶墉（1995）。**現代測驗理論**。台北：心理出版社。

李茂能（2000）。**中文電腦化適性測驗系統之應用與評鑑**。文景書局。

余民寧（1992）。試題反應的介紹-測驗理論的發展趨勢（二）。**研習資訊，9（1）**，5-9。

余民寧（1997）。**教育測驗與評量**：成就測驗與教學評量。台北: 心理。

洪碧霞、吳裕益、吳鐵雄、陳英豪（1992）。**能　估計方法、題庫特質及終止標準對CAT考生能　估計影響之研究**（國科會專題研究計畫成果報告編號：NSC81-0301-H024-03）。台北：中華民國行政院國家科學委員會。

洪碧霞、林素微、林娟如（2006）。認知複雜度分析架構對TASA-MAT六年級線上測驗試題難度的解釋力。**教育研究與發展期刊，2（4）**，69-86。

陳昇座（2007）。**以能　分佈為基礎之SHC曝光　控管法**。碩士論文未出版，國立臺中教育大學教育測驗統計研究所，台中市。

陳柏熹、王文中（2000）。**題間與題內多向　電腦化適性測驗**。發表於2000年教育與測驗學術研討年會。台北：台灣師範大學。

陳柏熹（2006）。能力估計方法對多向度電腦化適性測驗測量精準度的影響。**教育心理學報，38（2）**，195-211。

陳新豐（2004）。**線上題庫與適性測驗整合系統之發展研究**。發表於2004年教育及心理測驗學術研討會。台北：國立政治大學。

陳麗如（1998）。**電腦化適性測驗之題庫品質管理策略**。碩士論文未出版，國立臺灣師範大學資訊教育研究所，台北市。

張鈿富、王世英、吳慧子、周文菁（2006）。基本能力評量跨國發展經驗之比較。**教育資料與研究，68**，81-99。

黃吉楠（2004）。**多媒體英語文能力檢定暨適性化網路評量系統之建置**。碩士論文未出版，國立交通大學理學院網路學習碩士在職專班，新竹市。

楊曄齊（2006）。**以MOODLE 為平台之國中數學適性測驗工具**。碩士論文未出版，逢甲大學資訊工程所，台中市。

蔡文龍（2008）。**國小數學網路電腦化適性測驗系統之建置與研究—以國小三年級分數單元為例**。碩士論文未出版，嶺東科技大學數位媒體設計研究所，台中市。

錢永財、劉家惠、郭伯臣（2005）。**a-鄰近法選題對電腦適性測驗試題曝光率之比較**。發表於2005年教育與心理測驗學術研討會。台北：國立政治大學。

錢永財（2006）。**以a-鄰近法為選題策略之電腦化適性測驗模擬研究**。碩士論文未出版，國立臺中教育大學教育測驗統計研究所，台中市。

謝友詩、劉湘川、郭伯臣（2006）。電腦適性測驗題目曝光率之模擬研究。**測驗統計年刊，14（1），**62-74。

蕭顯勝、黃啓彥、游光昭（2006）。網路化科技素養適性測驗系統之建置。**理工研究學報，40（1），**1-21。

英文部份

Adams, R. J., Wilson, M. R., & Wang, W. (1997). The multidimensional random coefficients.

Andrich, D. (1982). An extension of the Rasch model for ratings providing both location and dispersion parameters. *Psychometrika, 47, 1*, 105-113.

Baker, F. B., & Kim, S. H. (2004). *Item Response Theory : Parameter Estimation Techniques*. Basel, N. Y. : Marcel Dekker, Inc.

Beguin, A. A., & Glas, C. A. W. (2001). MCMC estimation and some model-fit analysis of multidimensional IRT models. *Psychometrika, 66*, 541-561.

Birnbaum, A. (1968). Some latent trait model and their use in inferring an examinee’s ability. *In F. M. Lord and M. R. Novick, Statistical theories of mental test scores*, 17-20. Reading, Mass: Addison-Wesley.

Bock, R. D., Gibbons, R., Muraki, E. (1988). Full-information item factor analysis. *Applied Psychological Measurement*, 1988, 12, 261-280.

Bock, R. D., & Mislevy, R. J. (1982). Adaptive EAP estimation of ability in a microcomputer environment. *Applied Psychological Measurement, 6*, 431-444.

Carlson, J. E. (1987). Multidimensional item response theory estimation: a computer program. Unpublished manuscript.

Chang, H., Qian, J., & Ying, Z. (2001). a-stratified multistage computerized adaptive testing with b-blocking. *Applied Psychological Measurement, 25*, 333-341.

Chang, S. W., & Ansley, T. (2003). A comparative study of item exposure control methods in computerized adaptive testing. *Journal of Educational Measurement, 40*(1), 71-103.

de la Torre, J., & Song, H. (2009). Simultaneous Estimation of Overall and Domain Abilities: A Higher-Order IRT Model *Approach. Applied Psychological Measurement, 33*(8), 620-639.

Fraser, C. (1988). NOHARM. *[Computer software and manual]*. Armidale, New South Wales, Australia: author.

Hambleton, R. K., & Swaminathan, H. (1985). *Item response theory: Principles and applications*. Boston, MA: Kluwer-Nijhoff.

Hattie, J. (1981). *Decision criteria for determining unidimensional and multidimensional normal ogive models of latent trait theory*. Armidale, Australia: The University of New England, Center for Behavioral Studies.

Hung, P. H. (1988). *Application of computerized adaptive testing to the university entrance exam of* Taiwan, R. O. C. Unpublished doctoral dissertation, University of Minnesota, Minnesota.

Larson, J. W., & Smith, K. L. (1988). A Spanish computerized adaptive placement exam. UT: Brigham Young University Humanities Research Center.

Lee, J., Grigg, W., & Dion, G. (2007). *The Nation's Report Card: Mathematics 2007*. National Center for Education Statistics, Institute of Education Sciences, U.S. Department of Education, Washington, D.C.

Lord, F. M. (1977). Practical applications of item characteristic curve theory. *Jaurnal of Educational Measurement, 14*, 117-138.

Lord, F. M. (1980). *Applications of item response theory to practical testing problems*. Hillsdale, NJ: Lawrence Eribaum Associates.

Lord, F. M., & Novick, M. R. (1968). Statistical theories of mental test scores. Reading, MA: Addison-Wesley.

Masters, G. N. (1982). A Rasch model for partial credit scoring. *Psychometrika, 47*, 149-174.

McBride, J. R., & Martin, J. T. (1983). Reliability and validity of adaptive ability tests in a military setting. In D. J. Wiess (Ed.), *New Horizons in Testing: Latent Trait Test Theory and Computerized Adaptive Testing* (pp. 223-236). New York: Academic Press.

McDonald. R. P. (1982). Linear versus nonlinear models in item response theory. *Applied Psychological Measurement, 6*(4), 379-396.

McKinley, R. L. & Reckase, M. D. (1983). MAXLOG: A computer program for the estimation of the parameters of a multidimensional logistic model. *Behavior Research Methods and Instrumentation, 15*, 389-390.

Mislevy, R. J. (1991). Randomization-based inference about latent variable from complex samples.*Psychometrika,56*,Psychometric Society,Greensboro,pp.177-196.

Mislevy, R. J., A. E. Beaton, B. Kaplan and K. M. Sheehan.(1992). Estimating population characteristics form sparse matrix samples of item response. *Journal of Educational Measurement, 29*, pp.133-161, National Council on Measurement in Education,Washington, D.C..

Mullis, I.V.S., Martin, M. O., Ruddock, G. J., O'Sullivan, C.Y., Arora, A., & Eberber, E. (2005). TIMSS 2007 *Assessment Frameworks*. http://timss.bc.edu/TIMSS2007/frameworks.html.

Muraki, E. (1992). A generalized partial credit model: application of an EM algorithm. *Applied Psychological Measurement, 16*(2) ,159-176.

Muraki, E. (1999). Stepwise analysis of differential item functioning based on multiple-group partial credit model. *Journal of Educational Measurement, 36*, 217 – 232.

Muraki, E., & Carlson, E. (1995). Full-information factor analysis for polytomous item responses. *Applied Psychological Measurement, 19*, 73—90.

Muraki, E., & Bock, R. D. (1997). PARSCALE 3: IRT *based test scoring and item analysis for graded items and rating scales*. Chicago: Scientific Software International.

NAEP Technical Documentation (2009). *The Nation's Report Card. Retrieved* May 13, 2009, from National Center for Education Statistics: http://nces.ed.gov/nationsreportcard/tdw/.

OECD (2009). PISA 2006 *Technical Report*. OCED, Paris.

OECD (2005). PISA 2003 *Technical Report*. OCED, Paris.

Rasch, G. (1960). *Probabilistic Models for Some Intelligence and Attainment Tests*. Danish Institute for Educational Research, Copenhagen.

Reckase, M.D. (1997). A linear logistic multidimensional model for dichotomous item response data. In W.J. van der Linden & R. K. Hambleton (Eds.), Handbook of Modern Item Response Theory (pp. 271 – 286). NewYork: Springer-Verlag.

Reckase, M. D., & McKinley, R. L. (1991). The discriminating power of items that measure more than one dimension. *Applied Psychological Measurement, 15*, 361-373.

Ree, M. J. (1981). The effects of item calibrations, sample size, and item pool size on adaptive testing. *Applied Psychological Measurement, 5*, 11-19.

Segall, D. O. (1996). Multidimensional adaptive testing. *Psychometrika, 61*, 331-345.

Song, H. (2007). A higher-order item response model: development and application.Unpublished doctoral dissertation, The State University of New Jersey.

Stocking, M. L. (1994). *Three practical issues for modern adaptive testing item pools*. Educational Testing Service, Princeton, N. J.

Sympson, J. B. (1978). *A model for testing with the multidimensional items*. In D. J. Weiss (Ed.), Proceedings of the 1977 Computerized Adaptive Testing Conference (pp. 82-98). Minneapolis: University

of Minnesota, Department of Psychology, Psychometric Methods Program.

Thissen, D., Chen, W-H., & Bock, R. D. (2003). *Multilog (version 7)*. Lincolnwood, IL: Scientific Software International.

Veerkamp, W. J. J., & Berger, M. P. F. (1997). Some new item selection criteria for adaptive testing. *Journal of Educational and Behavioral Statistics, 22*, 203-226.

Wainer, H., Dorans, N. J., Flaugher, R., Green, B. F., Mislevy, R. J., Steinberg, L., & Thissen, D. (2000). *Computerized adaptive testing: A primer. 2nd edition*. Hillsdale, N.J.: Erlbaum.

Waller, N. G. (2002). WinMFact 2.0. Minneapolis, MN: Author.

Wang, T., & Vispoel, W. P. (1998). Properties of ability estimation methods in computerized adaptive testing. *Journal of Educational Measurement, 35*, 109-135.

Wang, W. C. (1994). Implementation and application of the multidimensional random coefficients multinomial logit model. Unpublished doctoral dissertation, University of California, Berkeley, CA.

Wang, W.-C., Chen, P.-H., & Cheng, Y.-Y. (2004). Improving measurement precision of test batteries using multidimensional item response models. *Psychological Methods, 9*, 116-136.

Wang, W. C., & Wilson, M. (2005). The Rasch testlet model. *Applied Psychological Measurement, 29(2)*, 126-149.

Weiss, D. J. (Ed.) (1983). New horizons in testing: Latent trait test theory and computerized adaptive testing. New York: Academic.

Weiss, D. J., & Yoes, M. E. (1991). *Item response theory*. In R. K. Hambleton and J. Zaal (eds.), Advances in educational and psychological testing. Boston: Kluwer Academic Publishers.

Wood, R., Wilson, D., Gibbons, R, Schilling, S., Muraki, E., & Bock, D. (2003). TESTFACT 4: Test scoring, item statistics, and item factor analysis. Mooresville, IN: Scientific Software.

Wright, B. D., & Masters, G. N. (1982). *Rating scale analysis*. Chicago: MESA Press.

Wright, B. D., & Stone, M. H. (1979). *Best test design. Chicago*: MESA Press.

Wu, M. L., Adams, R. J., & Wilson, M. R. (1998). Acer ConQuest. Melbourne, Victoria, Australia: Australian Council for Educational Research press.

Yao, L. (2003). BMIRT: Bayesian Multivariate Item Response Theory. [Computer software]. Monterey, CA: CTB/McGraw-Hill.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (1996). BILOG-MG: *Multiplegroup IRT analysis and test maintenance for binary items*. Chicago: Scientific Software International.