一、成就水準

NAGB 是以成就水準作為 NAEP 結果報告的主要方式，成就水準的設定就是要決定學生在不同的量尺分數點應該要學會什麼或會作什麼？在每個學科的每個年級中，定義三個水準為基礎（basic）、精熟（proficient）、進階（advanced）。定義這三個水準的程序主要簡述如下：首先配合測驗的內容和評量的技能，專家被要求定義出這三個水準操作性描述(operational descriptions);將這些描述記在腦海中之後，專家被要求評定哪些能力的學生會作對哪些題目且符合這些水準的操作性描述，最後將這些評定等級對應到 NAEP 的量尺中，得到成就水準的決斷分數。

二、試題圖的程序（**item mapping procedure**）

NEAP 設定二元計分試題答對率為 0.74、多點計分試題答對率為 0.65。Huynh（1998, 1994）指出二元計分試題（四個選項之試題）答對率在 0.75 時，該試題會有最大的訊息量。因此，設定受試者對於其能力分數鄰近之試題有 0.75 之答對率，並將估計之試題參數對照於量尺分數中。

三、評量架構

NEAP 評量架構包括學科內涵以及試題級數，以 NEAP 2007 數學為例，數學之學科內涵包括：「數字概念與運算、測量、幾何概念、分析與機率、代數」五項，而試題分為低階複雜、中階複雜、高階複雜三等級。NEAP 閱讀評量架構則分為「形成一般性的理解、發展解釋、讀者與文章之間的連結、檢視文章內容與架構」四層級。NEAP 科學的評量架構包含「地球科學、自然科學、生命科學」三領域，並評量學生「概念理解、科學探究、實際推理」三項科學關鍵能力。

第二節 PISA 大型測驗之探討

PISA是由OECD所委託的計畫，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的主要領域為閱讀，2003為數學，2006為科學。以下將簡要說明PISA實施時幾個重要之技術層面（OECD, 2005）。

壹、試題研發、測驗設計與背景問卷之發展

一、試題研發

PISA 試題研發過程包含初始準備（initial preparation）、審題會議（item paneling）、認知訪談（cognitive interview）、國際的審題會議、預試（pilot testing）。且為了讓考試工作能順利進行，有幾項工作需事先注意：（1）建立明確的施測流程；（2）受試者指導手冊；（3）監考人員指導手冊；（4）閱卷。

而在 PISA 認知試題的發展是由一套一系列廣泛的指導方針來引導，而這個指導方針在計劃開始時所擬定好的，並

且在 PISA2006 年科學專家小組第一次會議中所被認可的。而指導方針包含了發展的概要、試題需求的詳述。

在 PISA2000 與 2003 年是使用兩位數的編碼來區別，在每個試題必須要有對於反應的編碼，在每個編碼的原則包含了試題反應類別（包含全對、部分答對），在每個得分編碼都必須是不同的。

二、測驗設計

PISA 2003 評量以數學科為主，因此，測驗包含 7 個區塊的數學試題，M1~M7；2 個區塊的閱讀試題，R1 與 R2；2 個區塊的科學試題，S1 與 S2；2 個區塊的問題解決試題，PS1 與 PS2。每個試題區塊作答時間為 30 分鐘，則每個題本作答時間為 120 分鐘。

PISA 2006 年測驗試題包含 13 個試題區塊（7 個試題區塊為科學 S1-S7、2 個試題區塊為閱讀 R1、R2 與 4 個試題區塊為數學 M1-M4）。閱讀試題區塊（R1、R2）取自 2003 年之試題區塊，數學試題區塊（M1-M4）則為 2003 年之試題中挑選出 167 題試題組合而成，而 108 題科學認知試題中，有 22 題試題挑選自 2003 年，且分配至 7 個科學試題區塊中。

三、背景問卷

PISA 研發之背景變項問卷包含：學生問卷、學校問卷及提供參與國選擇的ICT（Information communication technology）熟悉問卷、父母問卷及全國性的問卷。所有問卷發展初期皆有經過預試的階段，一開始選擇澳洲先進行小樣本的抽測，讓學生對問卷內容進行自由討論，然後根據學生們的意見進行內容修訂，接著選擇日語系的日本、德語系的德國、法語系的加拿大及英語系的澳洲進行較大規模的預試，針對收集到的問卷預試資料進行分析，對學生們提出的問題或不適宜的題目進行增修刪補，以提高問卷試題的品質。

PISA 學生問卷大約需要花費學生 30 分中的填答時間，包含底下幾個面向的試題內容，學生特性：年級、年齡和性別...等；家庭背景：父母的職業、父母教育程度、家庭資源、家中藏書量，學生和父母的國籍，在家使用的語言...等；學生對於科學的看法；學生對於環境的看法；學生對於科學相關職業的看法；學習時間：包含在校及校外時間在不同科目課業上的學習模式與持續時間；學生對於接受科學教育的看法等等。

而學校問卷則提供給學校校長填答，約 20 分鐘可完成。內容涵蓋學校的組織架構、學校的人員及管理、學校資源、入學方式、科學及環境議題的教學、就業指導方面...等。另外 PISA 2006 有兩種問卷可提供參與國選擇，ICT（Information communication technology）熟悉問卷和父母問卷。ICT 熟悉問卷內容包含學生使用電腦的經驗、能力與頻率，以及對於使用電腦解決相關問題的自信等等的調查。而父母問卷內容則包含父母背景、子女的教育的花費、對環境的看法，以及對學校教育與科學教育的看法等等。

除此之外，參與國可以把全國性特殊問題增加到任何問卷，只是把全國性特殊問題插入到國際詢問表必須與國際研究中心達成協議，問卷作答時間不可設計超過 10 分鐘，且新增加的全國性問卷、ICT（Information communication technology）熟悉問卷和父母問卷於施測評量後都會被統一管理。

貳、抽樣設計與抽樣權重

PISA 目標母群為在所有參與施測國家中 15 歲的學生（大部分是七年級或是更高年級的學生），並使用二階段的分層抽樣設計，主要的抽樣步驟如下：

1. 定義各國的目標母群
2. 建立抽樣架構
3. 確認各抽樣層級（stratification）
4. 學校樣本的分配與挑選
5. 施測學生的挑選

PISA 使用二階段分層抽樣設計（two-stage stratified sample），第一階段是以學校為抽樣單位；第二階段是以學生為抽樣單位，針對該抽樣學校進行完全隨機抽樣。 由於在某一個施測國家內，就算對於學校或學生使用隨機抽樣進行樣本之選取，最終的施測樣本也不完全能代表全部的目標母群，因此，在進行資料分析時抽樣權重必須考慮。然而，由於每位施測樣本並沒有擁有相同被抽取機率，因此，PISA 在進行資料分析時必須考慮學校權重、學生權重、學校無作答反應之校正、年級無作答反應之校正、學生無作答反應之校正等因素。

參、測驗資料量尺化

PISA使用MRCML模式進行測驗資料分析，針對各項度之次級量尺進行估計，而使用軟體為ConQuest（Wu, Adams, & Wilson, 1997），多點計分試題使用PCM。個別受試者能力估計使用最大概似估計法（maximum likelihood estimation, MLE）估計受試者能力表現；群體能力估計使用可能值的方法。

一、可能值的分析

Mislevy 和 Sheehan（1987, 1980）根據插補理論（Rubin, 1987）提出可能值的概念，可能值是由量尺分數之邊際後驗分布中取出的隨機分數，且能合理地分配到每位受試者。可能值包含隨機誤差變異之組合，對於個人分數不是最佳的分數。但對於描述群體的表現時，可能值是一個較好的選擇（OECD, 2005）。

試題反應模式是一條件機率的模式，它描述了以能力值θ為條件而產生試題反應的過程。此模式完整的定義需要界定能力值θ的密度函數 $f_0(\theta;\alpha)$。令α為θ分佈的參數集。當定義單向度邊際試題反應模式（uni-dimensional marginal item

response models)，常假設抽樣的學生是來自於一個常態分布的母體，其平均數為μ，變異數為σ^2。也就是：

$$f_\theta(\theta;\alpha) \equiv f_\theta(\theta;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \qquad (2.2.1)$$

或者同義的式子，

$$\theta = \mu + E \qquad (2.2.2)$$

其中，$E \sim N(0,\sigma^2)$。

Adams、Wilson和 Wang. (1997)使用回歸模式$Y_n^T\beta$取代平均數μ，其中Y_n是一個u的向量，對於學生n，Y_n是固定且是已知，β是一個相對應的回歸係數向量。例如，Y_n可以由性別或社經水準等學生變項所構成。則學生n的母群模式可表示為

$$\theta_n = Y_n^T\beta + E_n \qquad (2.2.3)$$

其中，假設$E_n \overset{iid}{\sim} N(0,\sigma^2)$。

所以式子（2.2.1）可表示為

$$f_\theta(\theta_n; Y_n, b, \sigma^2) = (2\pi\sigma^2)^{-1/2} \exp\left[-\frac{1}{2\sigma^2}(\theta_n - Y_n^T\beta)^T(\theta_n - Y_n^T\beta)\right] \qquad (2.2.4)$$

這是一個平均數為$Y_n^T\beta$變異數為σ^2的常態分佈。如果式子（2.2.4）用來當作母群模式，則要估計的參數為β，σ^2及ξ。

如果是多維度變量母群模式，模式如下：

$$f_\theta(\theta_n; W_n, \gamma, \Sigma) = (2\pi)^{-d/2} |\Sigma|^{-1/2} \exp\left[-\frac{1}{2}(\theta_n - \gamma W_n)^T \Sigma^{-1}(\theta_n - \gamma W_n)\right] \qquad (2.2.5)$$

其中，γ是一個$u \times d$的回歸係數矩陣，Σ是一個$d \times d$的變異數共變數矩陣，W_n是一個$u \times 1$的固定變量向量。

在 PISA 中，W_n是條件變數(conditional variables)。結合條件機率的試題反應模式（式子 2.2.6）及母群模式（式子 2.2.5）可得到一邊際的試題反應模式(2.2.7)：

$$f(x;\xi \mid \theta) = \Psi(\theta,\xi)\exp[x'(B\theta + A\xi)] \qquad (2.2.6)$$

其中　$\Psi(\theta,\xi)=\{\sum_{x\in\Omega}\exp[z^T(B\theta+A\xi)]\}$

Ω：所有可能反應向量的集合

$$f_x(x;\xi,\gamma,\Sigma)=\int_{\theta} f_x(x;\xi\mid\theta)f_\theta(\theta;\gamma,\Sigma)d\theta \qquad (2.2.7)$$

在此模式下(2.2.7)，受試者的個別能力值是不被估計的。

每一位受試者的能力值之後驗分佈，如下所示：

$$h_\theta(\theta_n;W_n,\xi,\gamma,\Sigma\mid x_n)=\frac{f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;W_n,\gamma,\Sigma)}{f_x(x_n;W_n,\xi,\gamma,\Sigma)}=\frac{f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;W_n,\gamma,\Sigma)}{\int_{\theta} f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;W_n,\gamma,\Sigma)} \qquad (2.2.8)$$

在 PISA 中，式子 2.2.8 的模式使用在三個程序中:國家的校正(National calibrations)、國際間的校正(International scaling)、產生學生分數 (student score generation) 。

在國內的校正和國際間的量尺化時，條件試題反應模式(2.2.6)和母群模式(2.2.7)被使用，母群模式中並未使用到條件變數，也就是假設樣本是來自一多變量常態分布。

PISA2003的能力值包含七個向度:閱讀(Reading)、科學(Science)、問題解決(Problem solving)、數學(Mathematics)，其中數學又包含數量(quantity),空間和形狀(space and shape), 改變和關係(change and relationships) 不確定性(uncertainty)。當使用試題反應模式時，設計矩陣的設定如下:

設計矩陣:PCM (多元計分試題)、設計矩陣:Simple logistic model (二元計分試題)。

下面將簡述模式2.2.8如何使用於國家的校正、國際間的量尺化、產生學生分數。

國家的校正

國家的校正是使用未加權的資料(unweighted data)，每一個國家分開進行，校正的目的是要篩選和檢驗試題，主要有三種情況:

1.刪題:假如某一試題的特徵經過10個國家以上的分析都是不好的，則此試題會被刪除，此種試題又被稱為 "dodgy" item。

2.有些試題可能在某些國家中沒有被施測，因為這些試題的參數在這些國家分析的結果是不良，但在其他主要的國家這些試題卻表現良好

3.有些試題具有良好的參數特性，但卻也顯示試題和國家具有交互作用，即所謂的有差異性的試題，及試題的難度對於不同的國家而言是不同的。

上述第二類和第三類的試題都會對國家間的比較造成影響。

檢視國家的校正時會特別關注在試題對於量尺模式的適合度(the fit of the items to the scaling model)、試題鑑別度(item discrimination)、試題國家間的交互作用(item-by-country interaction)這三方面。

國際的校正

國際的試題參數的計算是利用模式2.2.6和模式2.2.7，同樣的在模式2.2.7中並未使用到條件變數。國際的校正樣本總共有15000學生，主要是從30個參與OECD的國家，每一個國家隨機抽樣500位學生而得。.

產生學生分數

在所有的試題反應模式中，學生的能力值是觀察不到的，它們是屬於遺失資料，需要從觀察得到的試題反應推論而得。有許多方法都可以推論能力值，PISA是使用多重插補的方式，也就是可能值。可能值是代表學生最有可能的能力值的值。下面將簡述可能值的使用。

使用國際間校正的試題參數，對於每一位學生，從能力值的邊際後驗機率(2.2.8)隨機抽取可能值。

PISA中，從模式2.2.8隨機抽取的步驟描述如下:

對於每一個受試者 n，M vector-valued random deviates, $\{\varphi_{mn}\}_{m=1}^{M}$,從多變量常態分佈，$f_\theta(\theta_n; W_n, \gamma, \Sigma)$ 。使用蒙地卡羅積分法逼近式子 2.2.8 的分母。

$$\int_\theta f_x(\mathbf{x};\xi\,|\,\theta) f_\theta(\mathbf{x},\gamma,\Sigma)\,\mathbf{d}\theta \approx \frac{\mathbf{1}}{M}\sum_{m=1}^{M} f_x(\mathbf{x};\xi\,|\,\varphi_{mn}) \equiv \Im \qquad (2.2.9)$$

同時，計算

$$P_{mn} = f_x(x_n;\xi\,|\,\varphi_{mn}) f_\theta(\varphi_{mn}; W_n, \gamma, \Sigma) \qquad (2.2.10)$$

$\{\varphi_{mn}, P_{mn}/\Im\}_{m=1}^{M}$的集合可視為式子 2.2.8 的後驗機率函數之近似；且機率值φ_{nj}可藉由以下公式求得：

$$q_{nj} = \frac{P_{mn}}{\sum_{m=1}^{M} P_{mn}} \qquad (2.2.11)$$

隨機產生 L 個服從均勻分佈的值$\{\eta_i\}_{i=1}^{L}$；對於每一次隨機抽取，若φ_{ni_0}滿足下列條件則選取當作一可能值向量(plausible vector)：

$$\sum_{s=1}^{i_0-1} q_{sn} < \eta_i < \sum_{s=1}^{i_0} q_{sn} \qquad (2.2.12)$$

建立條件變數

PISA 建立條件變數的方式主要是參考National Assessment of Educational Progress (NAEP) (Beaton, 1987) 和 TIMSS (Macaskill,Adams and Wu, 1998)。包刮下列幾個步驟:

步驟一：五個變數 (題本ID(booklet ID)、性別、母親的職業、父親的職業 和學校的數學平均分數) 直接視為是條件變數 。

步驟二：將學生問卷中的變數虛擬編碼(dummy coded)。詳見附件10。

步驟三: 對於每一個國家, 使用主成分分析分析虛擬編碼的變數並且計算每一位學生的主成分分數(主成份的數量必須要能解釋原始資料95%的變異才可以)。

步驟四：試題反應模式對於每一個國家的資料集是合適的且使用國際間校正的定錨試題的參數和經由主成分分析得到的條件變數估計國家的母群參數分佈

步驟五：使用上述的方法抽取五個可能值向量，每一向量的長度是7，代表7個PISA 2003所報告的能力值。

在PISA 2000中，如果學生沒有做到某一領域的任一題試題，則該位學生在該領域的可能值會被刪除而對於比較小的資料集則使用加權調整的方式，這種取向的假設沒有得到某一領域分數的學生資料是隨機遺失資料。但在PISA 2003中，所有學生在所有領域(domains)的可能值都被保留，這樣作有幾點好處：

1.因為不需要作加權調整，資料結構比較簡單且易於分析。

2.隨機遺失的假設可以得到一點鬆綁。產生可能值的假設是沒有任何試題反應被觀察到的領域和其他變數(條件變數和其他領域)的關係對於這兩群學生(有作到該領域的試題和沒有做到該領域試題的學生)是一樣的。使用所有這種關係訊息和所有關於學生的訊息插補學生的資料。因為關於資料的所有訊息都拿來協助插補資料，透過完整的資料集，我們將可以得到更準確的分析結果。再者，因為抽樣變異，有作答某一領域試題的學生特性和完全沒有作答該領域的學生特性應是相差不大的，而這樣的差異將在插補和估計學生特性的過程中被校正。舉個例子，針對所有學生所估計作閱讀能力的母群分布跟只針對實際有作閱讀領域試題的學生所作的閱讀能力分佈的估計應是差不多。

這種方法唯一的一個缺點是參照題本(PISA是booklet 9)的平均能力值將會影響那一些完全美有作到某一領域試題的學生的插補。假如某一個國家在參照題本中的某一個領域的能力值特別高或特別低，這種不尋常的表現將會影響完全沒有作到該領域試題學生資料的插補。

可能值的資料分析

可能值不是測驗分數，它們是從邊際後驗機率中隨機抽取出來可以合理代表個別受試者能力的值，因此可能值包含隨機誤差成分並不是個別受試者能力的最佳估計值，可能值是合用來描述母群的表現。我們可以使用標準的統計分析軟體，像是 SPSS 和 SAS，將可能值視為中介變項而得到母群參數的一致性估計的值，也可以使用ConQuest (Wu et al., 1997a)直接完成計算。

在PISA的學生檔案中包含40個可能值：

PV1MATH to PV5MATH ：數學素養mathematical literacy;

PV1SCIE to PV5SCIE ：科學素養 scientific literacy,

PV1READ to PV5READ ：閱讀素養 reading literacy and

PV1PROB to PV5PROB ：問題解決problem solving.

PV1MATH1 to PV5MATH1：數量quantity,

PV1MATH2 to PV5MATH2：空間和形狀space and shape

PV1MATH3 to PV5MATH3：改變和關係change and relationship

PV1MATH4 to PV5MATH4：不確定性uncertainty

$r(\theta,Y)$：每一位學生的能力值和可觀察變數的統計量，即 $(\theta,Y)=(\theta_1,y_{1,}\theta_2,y_2,\theta_3,y_{3,}\cdots\theta_N,y_{N,})$

(θ_n,y_n)：學生n的能力值和可觀察變數的值

θ_n 是觀察不到的，但我們可以觀察到作答反應 X_n

假如 $h_\theta(\theta;Y,\xi,\gamma,\Sigma\mid\mathbf{X})$ 是學生 n=1,2,....N的聯合後驗分佈函數，則我們可以藉由下列的式子計算 $r(\theta,Y)$ 的近似值

$$\gamma^*(X,Y)=E(\gamma^*(\theta,Y)\mid\mathbf{X},\mathbf{Y})$$
$$=\int_\theta\gamma(\theta,Y)h_\theta(\theta;Y,\xi,\gamma,\Sigma\mid\mathbf{X})d_\theta \qquad (2.2.13)$$

二、發展共同量尺

為比較PISA 2000與PISA 2003不同領域之表現，必須藉由定錨試題連結這兩年的分數量尺，包含（1）PISA 2000、PISA 2003閱讀素養與自然素養之量尺連結；（2）PISA 2000、PISA 2003數學素養之量尺連結。

其中2003年與2006年閱讀素養的可能值被量尺化到PISA2000年的量尺上，因為PISA2003年與2006年使用相同的試題，並使試題參數的估計在平均數為0，其中等化後線性轉換的方法與PISA2003年相同。在數學素養上，PISA2006年可能值被等化到PISA2003年的量尺上。另外在PISA2006年科學素養量尺上是另外建立一個全新的量尺，並沒有將PISA2006年進行線性轉換到與PISA2000年、2006年同一量尺。

肆、信度研究

測驗信度的檢測乃是測驗評量中重要的一環，PISA 針對 5 個量尺：數學、閱讀、科學、學習興趣與學習自信，使用可能值與 WLEs 之分析方式進行信度檢測，結果發現數學與閱讀之數據呈現 WLEs 法之信度較高，其餘三者以可能值分析法較高，但國際性的試題信度檢測皆在 0.8 以上。另外 PISA 針對 CR 試題提供三個評估信度的觀點，分別為同質性分析（homogeneity analysis）、變異數成分分析（variance component analyses）、各國之間的信度研究（inter-country

reliability study)，藉以評估各國間評分者一致性概況。而問卷背景變項之信度分析則以樣本加權過後之 Cronbach's alpha 值與驗證性因素分析(CFA)之結果為信度指標參考依據。

第三節 TIMSS 大型測驗之探討

TIMSS 主要目的為進行學生數學與科學教育成就趨勢調查研究，測試對象為 4 年級與 8 年級之學生，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 計畫每隔四年辦理國際數學與科學教育成就研究一次，並改名為 TIMSS。以下將簡要說明 TIMSS 實施時幾個重要之技術層面（Martin, Mullis, & Chrostowski, 2004）。

壹、評量架構、測驗設計與問卷之發展

一、評量架構

TIMSS 施測數學與科學兩學科，各學科的基礎架構由內容領域（content domain）與認知領域（cognitive domain）組成。TIMSS 2007 數學四年級的內容領域包含數（number）、幾何圖形與測量（geometric shapes and measures）、資料呈現（data pisplay），八年級內容領域包含數、代數（algebra）、幾何（geometry）、資料與可能性（data and chance）；認知領域則包含瞭解（knowing）、應用（applying）與推論（reasoning）。TIMSS 2007 科學四年級的內容領域包含生活科學（life science）、自然科學（physical science）、地球科學（earth science），八年級內容領域包含生物（biology）、化學（chemistry）、物理（physics）、地球科學（earth science）；認知領域則包含瞭解（knowing）、應用（applying）與推論（reasoning）。

二、測驗設計

TIMSS 2003 四年級測驗包含 313 題試題，其中，161 題數學試題與 152 題科學試題；八年級測驗包含 383 題試題，其中，194 題數學試題與 189 題科學試題。