第二章　文獻探討

為了達到本計畫之目的，本章簡要探討NAEP、TIMSS及PISA實施時幾個重要之技術層面，細節部份請參閱第四章研究成果。

第一節 NAEP 測驗之探討

NAEP為美國教育測驗服務社（Educational Testing Service, ETS）所發展的聯邦補助計畫，主要目的為建立學生學習成就的趨勢。NAEP是美國評量學生成就之代表，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）。NAEP評量之範圍可分為全國性的（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之評量分為主要評量（Main NAEP）與長期發展趨勢評量（Long-term Trend NAEP）兩類，主要的目的為（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）。以下將簡要說明NAEP實施時幾個重要之技術層面（Allen, Donoghue, & Schoeps, 2001）。

壹、測驗目標與評量架構之發展

一、測驗目標

NAEP主要可以分為全國性的（National）、各州的(State)、城市地區的（Urban District），主要目的是探索美國學生在主要的課程領域需要知道與具備的能力，並且長時間測量美國教育的發展情形。而為了達到這些目的，NAEP在計畫中包括了兩種重要評量類型，其中一個為主要評量（Main NAEP）與長期發展趨勢評量（Long-term Trend NAEP）。

NAEP 1998測驗科目包含閱讀、寫作及公民（civics），各科目之評量架構整理如下：

（一）閱讀能力之評量架構：為文藝學識而閱讀（reading for literary experience）、為獲得訊息而閱讀（reading to gain information）、為執行任務而閱讀（reading to perform a task）。

（二）寫作：說明文（informative）、記敘文（narrative）、議論文（persuasive）。

（三）公民：公民生活與政治學（civic life, politics, and government）、美國的政治體制的原則（the foundations of the American political system）、法規與美國體制（the constitution and American government）、美國與世界事務（the United States and world affairs）、美國公民的任務（the roles of United States citizens）。

二、測驗設計

NAEP 公民評量使用平衡不完全區塊設計（balanced incomplete block deign, BIB），而閱讀與寫作評量使用部份平衡不完全區塊設計（partially balanced incomplete block deign, PBIB），並且題本與受試者的配置上皆採取螺旋式（spiraling）分配。

BIB 設計與 PBIB 設計是將試題區分為數個區塊，將這些區塊有條件的編製成題本，且讓學生施測不同題本，以確保學生能夠接受並非完全相同的區塊題目，如此可確保不會有試題效應的產生，螺旋式分配指的是題本分配給受試者時採取螺旋式分配，以確保各測驗題本會有接近相同數量之受試者，如此可以確保後續分析參數的正確性能較佳。

三、教師問卷

NAEP 教師問卷對 4 年級及 8 年級教授閱讀、寫作及公民的教師進行調查。內容分為兩個部分，第一部份是有關於教師的背景和訓練，第二部份則是關於特定班級或單一班級教師的教學過程。在資料分析部分，教師問卷調查資料必須搭配每位教師所教授的所有受測學生之學習成就表現進行比對分析，學生可能被比對到教師問卷的第一與第二部分，對這些學生來說，問卷資料呈現他們教師的背景、訓練及對特定班級的特別教學方式。但大部分的學生只被比對到教師問卷的第一部分，畢竟特殊班級的學生樣本數量仍屬少數。

貳、抽樣設計（**sample design**）與抽樣權重 （**sampling weights**）

NAEP 施測樣本包含主要評量與長期發展趨勢評量的受測樣本，抽樣設計亦分為全國性評量的抽樣與州評量的抽樣。大致而言，施測樣本的選取包含以下幾個步驟：

1. 確認抽樣之目標母群與抽樣架構
2. 定義地理區域內主要的抽樣單位（primary sampling units, PSUs）
3. 由 PSUs 內挑選施測學校
4. 由施測學校內分配施測樣本的類型（包含一般受試者、殘障的受試者、英文不佳的受試者）與施測年級（4 年級、8 年級、12 年級在籍的學生）
5. 依據各年級挑選施測樣本

NAEP 使用多階段分層抽樣設計（multistage stratified cluster sample design）進行上述施測樣本的選取，主要抽樣分為四個階段：第一階段的抽樣單位是郡（PSUs）、第二階段的抽樣單位是小學與中學的學校、第三階段為抽樣學校之考科類型與樣本類型分配、第四階段為學生的挑選與考科類型的分配。由於 NAEP 進行抽樣時使用不同的抽取樣本比例，以及為了提高某些子群體特徵估計的準確性，進行超取樣（oversampling）來確保獲得較大的受試樣本。使得不同子群之施測樣本有不同被選取的機率，因此，每位受試者進行資料分析時，需確保每位受試者皆分配到一個權重。NAEP 權重是根據抽樣設計與反應不同類別個體的適當比例表現，這些權重程序包含：計算一個學生的基本權重（base weight）、完成不同年級的無作答反應調整（non-response adjustment）、整理（trimming）極度大的權重值、以及透過事後分層加權程序（poststratification procedures）來減少抽樣誤差等程序。

參、測驗信度（**reliability**）

NAEP 評估測驗信度指標包含：完全一致性百分比（percentage of exact agreement）、組內相關（the intraclass correlation）、Cohen's Kappa（Cohen, 1968）、積差相關係數（product moment correlation）等等。各項指標互有利弊，提供不同的分析情況下使用，在 C-R 試題分析上使用完全一致性百分比指標，在二元計分 C-R 試題使用 Cohen's Kappa 一致性指標，在多元計分 C-R 試題則使用組內相關指標作為參考依據。

肆、試題特性分析

NAEP 試題特性分析是採取各年級各學科領域分開進行分析的模式，分析項目包含背景試題與認知試題（二元計分試題與多點計分試題）兩部份。其中，二元計分試題分析使用標準化程序，試題的結果報告呈現數據包含：試題中各選項選答與遺漏樣本數之描述、受試者作答試題的百分比、受試者於該試題的答對率（p^+）、試題與答對分數的二系列相關係數（the biserial correlation coefficient）及點二系列相關係數（the point-biserial correlation coefficient）。而多點計分試題之結果報告則呈現包含：受試者作答試題的百分比、以試題之平均得分取代答對率、連續相關係數取代二系列相關係數、Pearson 相關係數取代點二系列相關係數等數據。

伍、**DIF（differential item functioning）分析**

DIF 差異試題功能分析之目的在提供一個控制群體間（between-group）差異之準則（測驗分數），觀察試題在不同群體受試者中是否有不同的難易度。藉由比較每一個試題在群體間的學習成就表現來獲知是否存在差異的訊息，NAEP 進行 DIF 分析是為了管理閱讀、寫作及公民能力測驗之試題，使試題品質更穩定。NAEP 比較三種參照群體與焦點群體：男生與女生、白人與黑人、白人與西班牙人。分析方法上採用以下三者：（1）M-H 方法（Mantel-Haenszel procedure, Mantel & Haenszel, 1959），（2）SIBTEST 方法（SIBTEST procedure, Shealy & Stout, 1993），（3）標準化方法（standardization method, Dorans & Kulick, 1986），詳細內容描述於本報告第四章第三節之 DIF 分析比較部份。

陸、量尺化程序

以 IRT 為基礎的測驗設計架構下，每個受試者被施予足夠的題數(60 題或 60 題以上)後透過估計的方法，如最大概似估計法(maximum likelihood estimate)可以準確估計個體的能力值 $\hat{\theta}$，此時個體能力值的測量誤差較小，可以忽略，能力值的分佈亦可以透過 $\hat{\theta}$ 近似而得。但是當測驗的內容廣泛且施測時間有限時，此時受試者只能被施予較少的題數測驗，則上述的優勢將不存在，即估計個體能力值的測量誤差將會變大而無法忽略，透過 $\hat{\theta}$ 的分佈近似母群的分佈將會產生極大的

偏誤(Wingersky, Kaplan,& Beaton, 1987)。即使題數足夠，但若是受試者所接受的測驗的形式不一樣，如題數、題型、試題的內容不相同，上述的問題一樣會發生。可能值方法則可以對群體參數提供一致性的估計結果，此基本方法已用於 1998 年的 NAEP 量尺分數報告，NAGB 提供了一個成就水準來判定量尺的意義。量尺化方法是指受試者於一個學科領域之表現，此表現是指受試者量尺分數或次級量尺分數。各學科領域之量尺是以 IRT 為基礎，並使用多重插補法（multiple imputation），即可能值的方法論估計量尺分數分布的特徵。基本的分析步驟概述如下：

步驟 1：使用 BILOG/PARSCALE 軟體估計參數。其中，BILOG 軟體用來估計混合 2PL(two parameter logistic model, 2PL model)與 3PL 模式(three parameter logistic model, 3PL model）的二元計分試題；PARSCALE 軟體用來估計 GPCM（generalized partial credit model, GPCM; Muraki, 1992）的多點計分試題；

步驟 2：依據已估計之試題參數，使用 MGROUP 軟體估計受試者之預測量尺分數分布（predictive scale score distributions）；

步驟 3：由預測量尺分數分布中隨機抽取計算統計特徵，例如受試群體之平均能力值；

步驟 4:決定適當的公制以建立量尺轉換機制，包含量尺之間的連結(linking)與轉換；

步驟 5：使用 jackknife 程序來估計不同群體中平均能力值的標準誤。

柒、學生表現描述

NAEP主要目標是告知社會大眾學生在學校內學了什麼與能做什麼的訊息，然而，NAEP量尺分數雖能提供不同子群體量尺分數的訊息，卻不能直接說明量尺上不同分數點所代表的涵義。傳統上，教育量尺的意義是附屬於常模參照上，而NAEP提出之成就水準與量尺分數點之描述是依據能力分類較可能表現出學生之分數水準，所以NAEP將試題對應到量尺分數點上，使得試題內容能提供學生會什麼的訊息。這種成就水準(achievement levels)的設定可見於1990年的數學測驗、1992年的閱讀測驗、1994年的歷史和地理測驗、1996年的科學測驗、1998年的寫作和公民測驗。

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。

TIMSS 2007 測驗試題四年級 353 題、八年級 429 題，各別分配至 28 個試題區塊，其中 14 個區塊為數學（M01-M14），14 個區塊為科學（S01-S14）（各區塊內僅包含數學或是科學單一領域題目），四年級與八年級之單數區塊（M01、M03...M13；S01、S03...S13）為由 TIMSS 2003 年挑選出之定錨試題區塊。

三、背景問卷

TIMSS 問卷分為四種類型，考科問卷：包含參與國四年級與八年級關於數學及科學課程的主題；學校問卷：學生的校長提供關於學校背景的資訊與關於數學和科學的教學資源；教師問卷：關於教師的背景，準備和專業訓練等，也詢問關於教學的活動，並收集詳細的教學訊息，此乃因為學生四年級時數學及科學通常是同一位老師教授，而八年級則為不同老師教授所設計；還有學生問卷：包含學生在校生活與在家學習數學與科學的經驗。他們被有系統的整合在 TIMSS2007 之課程模式中，此模式包含三個面向，預期、執行與獲得，也就是預期學生該學會的數學與科學課程內容；老師該教授的相關知識，包含如何教授與該由誰教授等等；以及學生已經學會什麼樣的課程內容或知識三個部分。

貳、抽樣設計與抽樣權重

TIMSS 的目標母群是指各國提供施測的母群體，主要是由兩個目標母群中挑選施測樣本，各國可以自由參加其中一個群體，或者是兩個都參加，其中，兩個母群體分為 4 年級（9 歲）與 8 年級（13 歲）在學的學生。此外，目標母群排除之樣本包含：智力有缺陷的學生、功能上（functionally）有缺陷的學生、以及非母語說話的學生。

TIMSS 使用多階段分層之集群抽樣設計（multistage stratified cluster design），其中，第一階段進行學校樣本的分層抽樣，第二階段則根據抽樣學校進行施測班級的抽樣。由於各國之受試者被抽測到的機率不同，因此，對於每位受試者必須計算其抽樣權重，抽樣權重的計算根據三個階段程序選擇不同的機率，包含學校、班級、以及學生。

參、試題分析

TIMSS2007之試題特性分析部份與TIMSS2003方法類似，皆為診斷性評量，估計所有施測試題的心理計量測量學上的參數，使用 IRT 試題反應理論。包含描述試題基本之參數估計，不同類型之信度分析，以及整合全部試題之分析內容。數學與科學試題包含選擇題及開放性試題，而開放性試題又分為二元計分試題與多點計分試題（0、1、2 三點計分），也就是填充題與應用題，其中，選擇題與二元計分試題分析採用 2PL 與 3PL 之 IRT 模式，多點計分試題則使用 GPCM；然而，進行量尺化程序前，測驗試題需進行簡單的描述性統計分析，包含整體測驗之統計描述、試題在各國之間之影響、測驗資料之信度研究等等。

肆、量尺化程序

藉由增加測驗的題數可以減少測量誤差，因此成就測驗時，題數常超過 70 題以獲取足夠的訊息，如此一來，伴隨每一θ的不確定性就可以被忽略，則θ的分布或是θ和其他變數的聯合分布就可以使用所估計θ近似而得。

當母群很大時，可以使用矩陣抽樣設計(matrix-sampling design)更有效率估計母群的能力分布，像是 TIMSS 所使用的。所謂矩陣抽樣設計:測驗內容範圍廣泛，每一位抽樣到的學生僅需做答部份測驗內容，當所有學生的答題反應被收集集合之後，可涵蓋所有的測驗內容。然而在這樣的設計之下，將無法準確的估計個體的能力，則上述的優勢將會無法存在，也就是個體能力的估計的不確定性將會太大而無法忽略，在這種其況下，集合個體的能力值估計母群的特性將會產生嚴重的偏誤(Wingersky,Kaplan,&Beaton,1987)。

可能值是解決此一問題之一方法，沒有先估計個體的能力然後再計算母群參數，可能值使用所有可得的資料，包含學生的答題反應和背景變項資料直接估計母群和次群體的參數。可能值是從估計的能力分布抽取而來，可以用在標準的統計分析軟體

1.可能值方法簡介

y：所有抽樣學生背景資料的反應

θ: 預估計的能力

假如所有抽樣的學生θ是知道的，則可以計算統計量$t(\theta, y)$，如樣本平均數或樣本百分點，而後推論相對應的母群參數T，可惜的是θ是未知的。將θ視為遺失資料並且用條件期望值近似$t(\theta, y)$。

給予學生的答題反應x_j，學生背景變數y_j，試題參數，從能力值的條件分布中隨機抽樣(可能值)可以近似t^*，計算t的θ值是從學生的條件分布中重複隨機抽取，Rubin(1987)指出這種重複的歷程可以將插補的不確定性量化，如透過不同的可能值集合，可以計算不同的t，這些t的平均，就是t^*的數值近似，他們所呈現的變異，反應無法直接觀察θ的不確定性。需注意的是，這種變異並未包含抽樣的變異，抽樣的變異藉由 jackknife variance estimation procedure 估計而得。

可能值並非估計學生的個別分數，而是對相似的學生(學生有相似的答題反應和背景變項)插補分數，這樣估計母群時會較準確。當模式被正確介定時，可能值可以提供母群參數的一致性估計，但他們並非個體能力的不偏估計，使用可能值的平均並不能代表個別學生的能力 Mislevey, Beaton, Kaplan, & Sheehan (1992)。

每一個學生 j 的可能值從條件分佈$P(\theta_j \mid x_j, y_j, \Gamma, \Sigma)$抽取

Γ：背景變數的回歸係數矩陣

Σ：殘差共變異矩陣

$$P(\theta_j \mid x_j, y_j, \Gamma, \Sigma) \propto P(x_j \mid \theta_j, y_j, \Gamma, \Sigma) P(\theta_j \mid y_j, \Gamma, \Sigma) = P(x_j \mid \theta_j) P(\theta_j \mid y_j, \Gamma, \Sigma)$$

$P(x_j \mid \theta_j))$：試題反應模式

$P(\theta_j \mid y_j, \Gamma, \Sigma)$：在背景變項$y_j$、參數$\Gamma$和$\Sigma$的條件下，能力值的多變量聯合密度函數。在計算的過程中，試題參數是固定的並且被視為是母群的值。

2.條件(Conditioning)

$P(\theta_j \mid y_j, \Gamma, \Sigma)$被假設為是一多變量常態分布，共變異數是$\Sigma$，平均數是迴歸參數$\Gamma$的線性模式。在 TIMSS 中使用 PCA 減少背景變數的個數然後使用在Γ中可以解釋原始資料 90%的變異的成分被使用，這些成分就是條件變數，以y^c表示，模式如下:

$\theta = \Gamma' y^c + \varepsilon$

ε 是常態分布，平均數是 0，變異數是 Σ

Γ 是一矩陣每一欄是每一個能力量尺的效果(effects)

Σ 是量尺之間的殘差變異矩陣。

為了要正確估計上述的函數 $\theta = \Gamma' y^c + \varepsilon$，對於所有的背景變數，$P(\theta \mid y)$ 需正確被界定。如果在估計包含條件變數的函數 Γ 時不是在此種情況下（$P(\theta \mid y)$ 需正確被界定），將會因為不正確的界定(misspecification)而產生誤差。

在 TIMSS2007，以幾乎所有背景變項為基礎的主成分分數被使用。這些背景變項高度反應教育政策和教育實務，透過這些變數所計算的 θ 的邊際平均和百分點幾乎是最佳的。

3.產生能力值(generating proficiency scores)

步驟一 :從一個近似常態的分配 $P(\Gamma, \Sigma \mid x_j, y_j)$，固定 Σ 為 $\hat{\Sigma}$，抽取一個 Γ。

步驟二:在 Γ 的條件下，（且固定 $\Sigma = \hat{\Sigma}$），公式 7 後驗分佈的平均 θ_j 和變異數 $\sum_j^p$ 使用 EM 的演算法則計算。

步驟三: 能力值從一個多變量常態分佈（平均 θ_j、變異數 $\sum_j^p$ ）獨立抽取。

這三個步驟重複五次，每一位學生產生 5 個 θ_j 的差補值。

學生們雖然被施測較少的題數，但是學生的 Γ 和 Σ 是固定的，因此所有的學生不管施測的題數都被指定一組可能值。

4.條件變數

(1)對於類別變項，每一個選項使用虛擬變項編碼，假如學生沒有作答（遺漏）或沒有被施測，那一題的虛擬編碼被設定是 0。

(2)連續變項的背景資料，像是出生年，家中人口數是使用效標量尺(criterion scaling)重新編碼。就是每一個反應選項使用 interim achievement score 代替。

(3)每一個國家，所有的虛擬編碼的變數和效標量尺(criterion-scaled)的變數被包含入主成分分析。這些主成分需能解釋背景變項 90%的變異。因為每一個國家的主成分分析是分開計算的，因此每一個國家的主成分成分個數可能不大一樣。

(4)除了主成分分析萃取的成分，性別(dummy-coded)、試卷使用的語言(dummy-coded)、學生所隸屬的學校班級(criterion-scales)、特定選擇的國家變數(dummy-coded)是主要的條件變數，如此一來，將能解釋最大的學生之間的變異並且保留教室之間和教室內的變異。

在TISSS2007技術報告中明確指出，要將IRT量尺化和可能值方法應用於TIMSS2007評量中有四個主要的工作：

1. 校準測驗試題（估計各個試題參數）
2. 在學生問卷的條件變數中找出主要成分
3. 建立數學與科學整體的 IRT 量尺（精熟分數）、數學與科學在各個內容與認知領域的 IRT 量尺（精熟分數）
4. 將量尺上的精熟分數與前一次測驗做比較

本研究主要目的為建立一套適合 TASA 之標準化流程，因此，首先就國外大型測驗（NAEP、TIMSS、PISA）進行相關文獻之整理與分析，同時探討各研究步驟之優缺點，以發展適用於 TASA 之標準化測驗。根據文獻探討，本計畫整理欲探討大型標準化測驗實施時之重要程序，主要針對以下部分：抽樣權重、測量模式、試題特性與背景變項分析、量尺化程序及結果報告之呈現。