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）。