9789860313840

大型標準化測驗建置流程應用於TASA之研究

國家教育研究院 編印

大型標準化測驗建置流程應用於TASA之研究

主編：郭伯臣、曾建銘、吳慧珉

國家教育研究院 101年1月編印

序

許多先進國家，對於學生基本能力表現相當關切，因此，持續進行教育資料庫的建置。近年來，國內也逐漸重視教育資料庫之建置，特別是在九年一貫課程實施之後，陸續有全國性學生學習成就資料庫之建置，例如：臺灣教育長期追蹤資料庫（Taiwan Education Panel Survey, TEPS）、臺灣高等教育資料庫、臺灣學生學習成就評量資料庫（Taiwan Assessment of Student Achievement, TASA）等，足見國內教育對此重視之端倪。

然而，國內教育資料庫大部分是以參照國外資料庫之建置方式，例如：試題研發、等化設計、施測流程、背景變項調查等，仍尚未制定一套符合國內教育資料庫之標準化測驗建置流程。除了使得資料庫僅侷限於國內學生間相互比較，無法達到與國外教育資料庫進行連結對照之功效外，亦無法將其價值與貢獻發揮至極致。

近幾年國內積極參加一些國際評比之大型測驗（large-scale assessments），如美國國家教育進展評量（National Assessment of Educational Progress, NAEP）由於其實施的時間較早（1968），成為許多大型測驗之參考依據；PISA（國際學生評量）測驗是由「經濟合作與發展組織（Organization for Economic Co-operation and Development, OECD）」主辦，自從1997年起分別進行四次跨國的學生評量測驗（PISA 2000、PISA 2003、 PISA 2006及PISA 2009），國內參與第三、四次跨國學生評量測驗（PISA 2006、PISA 2009）；另國內亦參與「國際成就調查委員會（The International Association for the Evaluation of Education Achievement, IEA）」主辦之「第三次國際數學與科學教育成就研究後續調查（Third International Mathematics and Science Study Repeat, TIMSS-R）」（之後改名為TIMSS）。除了參與測驗的學童，國內許多研究人員亦參與此大型調查研究之相關工作。

雖然國內教育系統已邁向國際化，進行與國際評比接軌之行動。然而，國內參與之國際評量，仍是以抽樣設計、題本設計、測驗編製與內容等前置作業為主；相對的，國內大型標準化測驗之建置，對

於較為核心之技術相對缺乏，例如：抽樣設計與抽樣權重（sampling weights）、測量模式、試題特性與背景變項資料分析、量尺化程序（scaling procedures）等，而這些方面國外大型測驗則提供了良好的模式。因此，本書藉由分析這些大型測驗（TIMSS、PISA、NAEP）的測驗發展程序與量尺建立之方式，期協助國內大型標準化測驗於建立流程時，能依循標準化程序，使資料庫更趨嚴謹且具信效度。

國家教育研究院院長

目錄

表次

圖次

第一章　緒論

國外許多先進國家的教育系統，對於學生基本能力表現都有相當深切的關懷及具體明確的認知，因此，這些國家持續地進行教育資料庫的建置。國內教育系統也漸漸重視教育資料庫之建置，特別是在九年一貫課程實施之後，陸續有全國性學生學習成就資料庫之建置計畫，例如：臺灣教育長期追蹤資料庫（Taiwan Education Panel Survey, TEPS）、臺灣高等教育資料庫之建置及相關議題之探討、臺灣學生學習成就評量資料庫（Taiwan Assessment of Student Achievement, TASA）等，足以見到國內教育對此重視之端倪，其中臺灣學生學習成就評量資料庫為結合國內大學院校、學術研究機構等學者專家之學術專長以及資深教師的經驗，建置臺灣地區國小四年級、國小六年級、國中二年級、高中二年級與高職二年級學生學科學習成就評量資料庫。其主要目的如下：

1. 建立國民中小學、高中及高職學生學習成就長期資料庫，以追蹤、分析學生在學習上變遷之趨勢，進而檢視目前課程與教學實施成效。
2. 提供完整、標準化的學習成就資料，作為分析學生學習成就上差異表現變項資料，以評估學生未來在學術方面能力之發展與社會期許。
3. 瞭解國內學校教學及學生學習成效之現況，作為課程與教學政策改進之參考，並為縣市政府教育局及學校推動補救教學之重要參據。
4. 提供各縣市學生學習表現資料，建立與縣市合作機制，以擴大資料庫應用效益。
5. 以資料庫的量化資料，提供國內外相關研究人員，深入探討學生學習成就方面的相關政策議題。
6. 建立本國學生學習成就評量資料庫，同時考慮與國際接軌，利於加入國際比較行列，藉以瞭解臺灣教育之獨特面與優缺點。

然而，國內教育資料庫主要是以仿造國外資料庫之建置方式，例如：試題研發、等化設計、施測流程、背景變項調查等，卻沒有制定一套符合國內教育資料庫之標準化測驗建置流程。除了使得資料庫僅侷限於國內學生間相互比較，無法達到與國外教育資料庫進行連結對照之功效外，亦無法將其價值與貢獻發揮至極致。

近幾年國內積極參加一些國際評比之大型測驗（large-scale assessments），其中國家教育進展評量（National Assessment of Educational Progress, NAEP）由於實施的時間較早，成為許多大型測驗之參考依據，這些大型測驗分別說明如下：

一、國家教育進展評量

美國全國教育統計中心（National Center for Education Statistics, NCES）的最高行政長官負責執行NAEP政策，並由全國評量管理委員會（National Assessment Governing Board, NAGB）所制定之政策指導下執行其功能。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、數學、科學等科目之評量（The Nation's Report Card, 2005）。2005 TASA英語科與數學科即是參考NAEP之評量架構，以檢測學生之能力。

二、國際學生評量（The Programme for International Student Assessment, 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，至從1997年起分別進行三次跨國的學生評量測驗（PISA 2000、PISA 2003及PISA 2006），國內參與第三次跨國學生評量測驗（PISA 2006），並積極參與PISA 2009前置作業。

三、國際數學與科學教育成就趨勢調查（Trends in International Mathematics and Science Study, TIMSS）

行政院國科會於1999年起參與「國際成就調查委員會（The International Association for the Evaluation of Education Achievement, IEA）」主辦之「第三次國際數學與科學教育成就研究後續調查（Third International Mathematics and Science Study Repeat, TIMSS-R）」，之後並改名為TIMSS。國內學童除參與TIMSS-R、TIMSS 2003及TIMSS 2007之測驗外，國內許多研究人員亦參與此大型調查研究之相關工作。

由上述可知，國內教育系統已邁向國際化，進行與國際評比接軌之動作。目前國內參與之國際評量，仍是以抽樣設計、題本設計、測驗編製與內容等前置作業為主，國內大型標準化測驗之建置，對於較為核心之技術仍較為缺乏，例如：抽樣設計與抽樣權重（sampling weights）、測量模式、試題特性與背景變項資料分析、量尺化程序（scaling procedures）等，這一方面國外大型測驗提供良好的範例，因此，本書欲藉由分析這些大型測驗（TIMSS、PISA、NAEP）的測驗發展程序與量尺建立之方式，深入探討如何使用標準化程序建立嚴謹之大型標準化測驗。

近年來，隨著資訊科技快速進步、測驗形式的改變及測量的概念日趨複雜，大型測驗之評量亦開始採用較複雜之測驗題型，例如：填充題、簡答題之類的建構反應試題（constructed response item），或是題組試題等，此類試題計分規則較為複雜且一份測驗或是一題試題可能測量許多不同的能力或特質，因此，必須配合適當的「測驗理論」才能從學生答題反應中萃取出所要瞭解的認知能力。

由PISA、NAEP、TIMSS技術報告公佈的評量架構，清楚呈現其測量之能力不單純的只有單一能力，這樣的測驗可能是多向度（OECD, 2006; The Nation's Report Card, 2009; Mullis, Martin, Ruddock, O`Sullivan, Arora, & Erberber, 2007）；然而，當試題是測量多向度能力，卻仍以單向度試題反應理論（unidimensional item response theory, UIRT）進行參數估計，將會產生偏差的試題參數估計和能力參數（Ackermean, 1991）。目前 NAEP、TIMSS仍以UIRT為主要使用之測量模式，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)表現較無法做精確描述；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：如果測驗為題組試題之測驗題型，但卻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。

在NAEP、TIMSS及PISA中，母群或母群中某些群體之能力表現為大家所關注之議題，國內常見的方式是直接使用個別受試者的成績（能力值）對母群或個別群體的表現進行估計，常以個別受試者的成績（能力值）平均值或變異數代表該群體之某一能力表現及其分散程度，更進一步進行各種假設檢定，例如： TASA數學科即採用此方式（洪碧霞、林素微、林娟如，2006）。依據Mislevey等人（Mislevey,1991; Mislevy, Beaton, Kaplan, & Sheehan, 1992; OECD, 2005; Lee, et al., 2007）之研究結果顯示，此種推論母群表現之方式容易造成偏誤。根據Mislevey等人之研究，可能值（plausible values）包含隨機誤差成分，不適合描述個體分數，但可能值具有良好群體估計一致性，適合描述群體之特性（Mislevey,1991; Mislevey, et al., 1992）。因此，目前國際上大型測驗皆以此種技術進行群體統計特性描述（OECD, 2005; Lee, et al., 2007），本書擬深入探討此一技術之使用方式，運用於臺灣學生學習成就評量資料庫（TASA），進而介紹國內相關研究領域使用。

綜合上述可知，目前國際上較知名的大型標準化測驗在評量架構、試題與測量模式之配合上仍有不一致與不足之處。因此，本書擬探討NAEP、TIMSS及PISA等測驗之資料分析步驟與方法，包含抽樣設計與抽樣權重、測量模式、試題特性與背景變項資料分析、量尺化程序等，進而提出適用於國內TASA之標準化資料分析步驟與方法。

第二章　文獻探討

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與長期發展趨勢評量。

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、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、差異試題功能分析（differential item functioning, DIF）

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），詳細內容描述於本書第四章第三節之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θ，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θ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θ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，而試題分為低階復雜、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（大部分是九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$f_\theta(\theta;\alpha)$。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：

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

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

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：

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

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的能力值包含四個向度：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，$f_{\theta}(\theta_{n};\mathrm{W_n},\Upsilon,\Sigma)$。使用蒙地卡羅積分法逼近式子2.2.8的分母。

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

同時，計算

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

$\{\varphi_{\mathrm{mn}}, P_{\mathrm{mn}}/\Im\}_{\mathrm{m=1}}^{M}$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：

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

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NAEP（Beaton, 1987）和 TIMSS（Macaskill, Adams and Wu, 1998）。包括下列幾個步驟：

步驟一：五個變數（題本ID（booklet ID）、性別、母親的職業、父親的職業 和學校的學科平均分數）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）。

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，資料結構比較間單且易於分析。
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沒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

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 $h_\theta(\theta; Y, \xi, \gamma, \sum | \mathbf{X})$ 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

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

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，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作答部份測驗內容，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，在這種情況下，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。

一、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 Γ 和 Σ 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。

二、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 Σ，平均數是迴歸參數 Γ 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。

三、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。

四、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。

本書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，試圖探討大型標準化測驗實施時之重要程序，主要則針對以下部分：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。

第三章　研究方法

本書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，透過研究、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來進行。

壹、研究步驟

一、文獻蒐集

針對本書所欲探討之國外大型測驗NAEP、TIMSS、PISA進行文獻之蒐集。

二、進行文獻探討

本書將針對國內外大型測驗NAEP、TIMSS、PISA以及TASA進行文獻探討，其中又分為抽樣權重、測量模式、試題特性與背景變項分析、量尺化程序、結果報告等五大部分。

三、建立適合TASA之標準化資料分析步驟與方法

透過文獻的探討，瞭解國外大型測驗之標準化資料分析步驟與方法，並建立適合TASA資料之標準化資料分析步驟與方法。

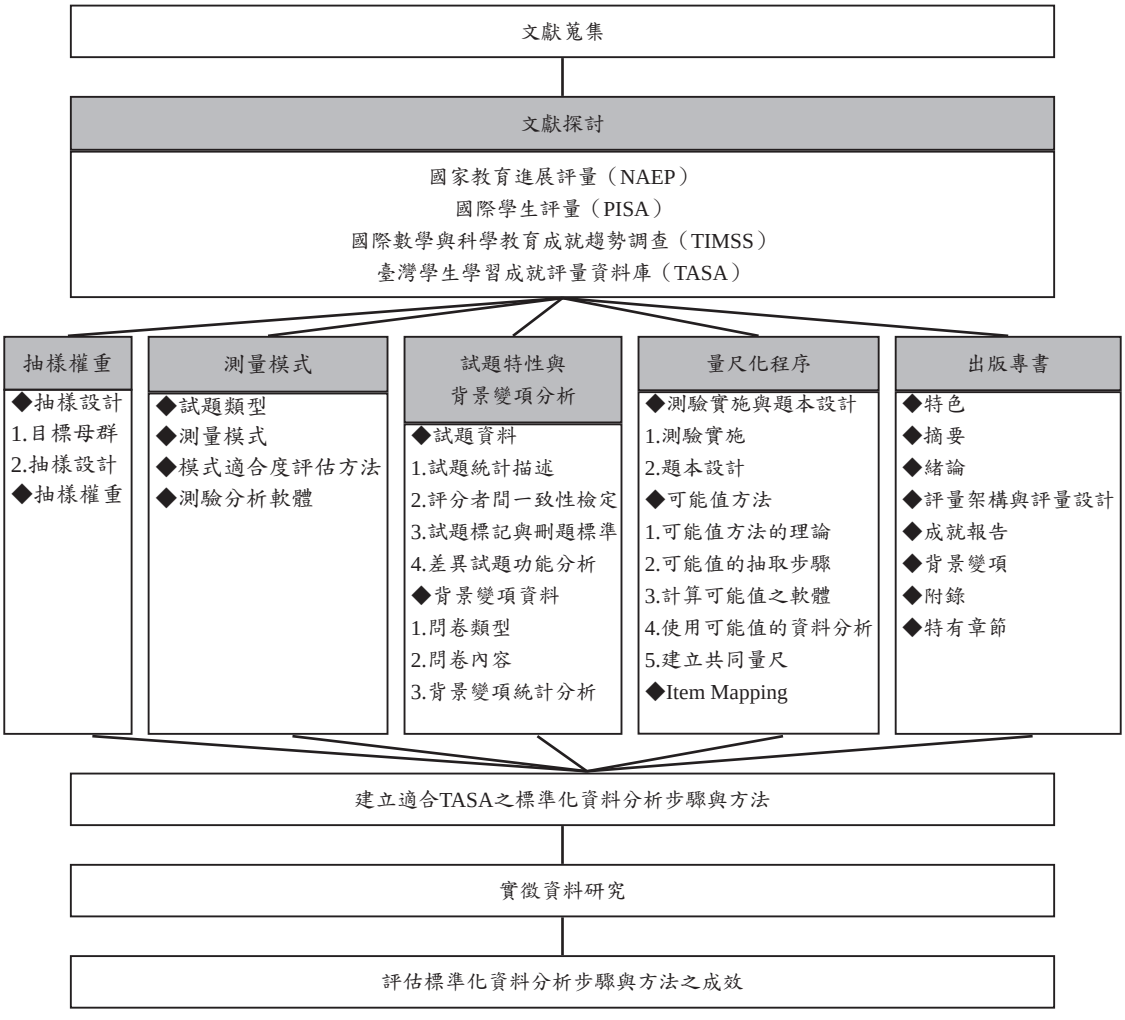
四、以TASA為實徵資料進行研究

本書的實徵資料以TASA 2006小四數學科測驗為例，透過所建立之標準化資料分析步驟與方法來進行該測驗資料的分析。

五、評估標準化資料分析步驟與方法之成效

透過實徵資料與模擬研究，用來評估所建立之標準化資料分析步驟與方法之成效，並提出適合TASA資料的分析步驟與方法。

貳、研究流程


文獻蒐集
文獻探討
國家教育進展評量（NAEP）
國際學生評量（PISA）
國際數學與科學教育成就趨勢調查（TIMSS）
臺灣學生學習成就評量資料庫（TASA）
抽樣權重
◆抽樣設計
1.目標母群
2.抽樣設計
◆抽樣權重
測量模式
◆試題類型
◆測量模式
◆模式適合度評估方法
◆測驗分析軟體
試題特性與背景變項分析
◆試題資料
1.試題統計描述
2.評分者間一致性檢定
3.試題標記與刪題標準
4.差異試題功能分析
◆背景變項資料
1.問卷類型
2.問卷內容
3.背景變項統計分析
量尺化程序
◆測驗實施與題本設計
1.測驗實施
2.題本設計
◆可能值方法
1.可能值方法的理論
2.可能值的抽取步驟
3.計算可能值之軟體
4.使用可能值的資料分析
5.建立共同量尺
◆Item Mapping
出版專書
◆特色
◆摘要
◆緒論
◆評量架構與評量設計
◆成就報告
◆背景變項
◆附錄
◆特有章節
建立適合TASA之標準化資料分析步驟與方法
實徵資料研究
評估標準化資料分析步驟與方法之成效

參、研究工具

一、MATLAB軟體

二、SPSS軟體

三、測驗資料分析之軟體

1. BILOG-MG（Zimowski, Muraki, Mislevy & Bock, 2003）
2. PARSCALE（Muraki & Bock, 1996）
3. ConQuest
4. SCORIGHT（Wang, Bradlow, & Wainer, 2004）

第四章　研究結果

第一節　抽樣設計與抽樣權重

抽樣（sampling）是由目標母群（target populations）中選取基本單位為樣本，並且認為能透過樣本選取進行目標母群特徵之推估。透過許多抽樣方法選取之樣本仍需依賴著母群特徵與調查研究的問題，因此，抽樣設計的選擇程序將以能避免偏誤產生與考慮在有效資源內達到最佳的精準度為主。然而，不論選擇哪一種抽樣設計仍然會有偏誤的產生，例如：假設使用非隨機（non-random）的方法進行抽樣，代表這個選擇是有意識地或無意識地受到人為的影響；或是假設抽樣架構（sampling frame）的選擇沒有足夠覆蓋整個母群、不夠完整或準確等。

抽樣權重被定義為樣本被抽取機率的倒數，而在調查研究中，受試樣本的抽取是透過不同組群進行選取，使得每位受試樣本被抽取之機率不同，因此，必須加上抽樣權重（sampling weights），以正確地推估目標母群（Martin & Kelly, 1996）。一般而言，若調查研究使用簡單隨機抽樣進行受試樣本之選取，則由於每位受試樣本被選取的機率相同，因此，每位被抽取的受試樣本抽樣權重值也會相同，且樣本權重總和為目標母群的個數。也就是說，若使用簡單隨機抽樣進行樣本之抽樣，則不需考慮到抽樣權重的問題。然而，在調查研究或大型測驗中卻很難使用簡單隨機抽樣的方法，主要原因包含（OECD，2005）：

1. 花費太高

   由於大型測驗的受試樣本是由學校母群中所抽取，若使用簡單隨機抽樣則受試者相當可能分布在許多不同學校之中，因此，必須花費相當多的費用。

2. 不實際

   除了費用太高以外，對於行政業務來說，必須聯絡與接觸相當多的學校，使得這樣的抽樣方法對於施測進行相當不實際。

3. 對於欲比較之變項，無法進行連結比較

   若使用簡單隨機抽樣，以統計的觀點來看，對於學生、學校、班級、老師等變數將無法進行有效的連結比較。由於在抽樣的過程中，學生或學校有可能只抽取到一個或少數幾個樣本，這樣的樣本數無法進行較穩定的統計推論。

綜合上述，可知簡單隨機抽樣方法很難使用於教育測量之中。因此，本書探討國外大型測驗（NAEP、PISA、TIMSS）的抽樣架構，檢視TASA現行的抽樣架構是否需要修正，以建議未來TASA將使用的抽樣設計。此外，大型測驗通常是透過多階段的抽樣方法抽取受試樣本，此方法將使得抽取之受試樣本並非在相同情況下被抽取，即受試樣本被抽取之機率不相等，因此，必須搭配使用正確的抽樣權重，才能正確地推論目標母群。以下將分別針對NAEP、PISA、TIMSS、TASA抽樣設計進行探討、介紹如何計算抽樣權重、抽樣變異估計方法、現行TASA抽樣設計的限制、以及提出適合TASA採用之抽樣設計。

壹、NAEP、PISA、TIMSS、TASA抽樣設計介紹

一、NAEP（Allen, Donoghue, & Schoeps, 2001）

NAEP全國抽樣設計是使用多階段抽樣設計（multistage probability sample），主要分為四個階段。第一階段是以郡（county）為抽樣單位、第二階段是以學校為抽樣單位、第三階段為考科與樣本類型的分配、第四階段為受試學生的選取與考科類型的分配（如圖4-1-1所示）。

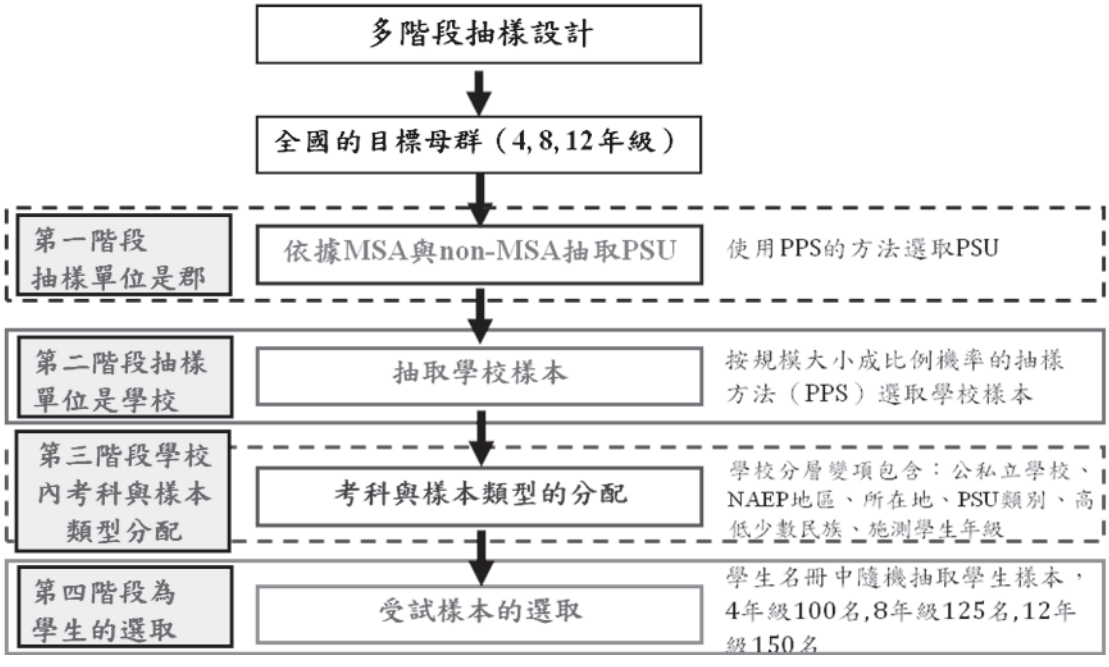


圖4-1-1　NAEP抽樣架構

本書以NAEP 1998技術報告內容進行介紹，NAEP目標母群為公私立學校四年級、八年級、十二年級入學的學生（小學與中學），施測樣本亦包含黑人與西班牙人的學生，以及殘障（students with disabilities, SD）與英文能力不佳的學生（limited English proficient, LEP）的學生（僅閱讀測驗）。預計抽測樣本如表4-1-1所示，可以發現考科施測人數的比例不同，例如：4年級寫作/公民26000人（佔72%）；公民特別趨勢2000人（佔6%）；閱讀8000人（佔22%），使得NAEP進行受試者考科分配時，即按照此比例進行分配。

表4-1-1　NAEP預計抽測學生數

	Subject		Target Sample Size
Total			**132,000**
Grade 4	Civics	6,000	
	Civics Special Trend	2,000	
	Reading	8,000	
	25-Minute Writing	20,000	
		Grade 4 Total	**36,000**
Grade 8	Civics	8,000	
	Civics Special Trend	2,000	
	Reading	11,000	
	25-Minute Writing	20,000	
	50-Minute Writing	6,000	
		Grade 8 Total	**47,000**
Grade 12	Civics	8,000	
	Civics Special Trend	2,000	
	Reading	13,000	
	25-Minute Writing	20,000	
	50-Minute Writing	6,000	
		Grade 12 Total	**49,000**

資料來源：NAEP 1998 Technical Report, p.33

表4-1-2　4年級閱讀評量等化設計（PBIB設計）

題本序號	區塊位置	區塊位置	題本序號	區塊位置	區塊位置
1	R4	R3	9	R7	R8
2	R3	R5	10	R8	R6
3	R5	R9	11	R6	R7
4	R9	R4	12	R10	R8
5	R4	R5	13	R7	R4
6	R3	R9	14	R8	R3
7	R6	R10	15	R5	R6
8	R10	R7	16	R9	R10

此外，每年級考科抽樣人數的設定，是由各考科試題區塊大約要有2000位受試者進行施測所推估。例如：表4-1-2為4年級閱讀評量使用的等化設計，此設計由16個測驗題本（1~16）與8個試題區塊（R3~R10）組合而成，每個測驗題本包含2個試題區塊。試題區塊至少會有四種組合（8/2），因此，為了符合每個試題區塊大約要有2000位受試者進行施測之條件，NAEP預計閱讀應抽取知樣本數為8000人。

NAEP 1998實際抽測樣本為：94個主要抽樣單位（primary sampling units, PSU），4年級共733間參與學校、8年級共761間參與學校、12年級共608間參與學校， 4年級共36104名學生參與施測、8年級共48797名學生參與施測、12年級共48588名學生參與施測，總計133489名學生。接著，分別介紹NAEP各階段抽樣方式。

第一階段（抽取PSU）

共1027個PSUs，其中PSU是由一個綜合都會統計區（consolidated metropolitan statistical area, CMSA）、都會統計區（metropolitan statistical area, MSA）、新英格蘭統計區（New England County metropolitan area, NECMA）、郡、或美國鄰近郡的群體（阿拉斯加州、夏威夷、哥倫比亞地區）等所構成。總計有290個是MSAs與737個為non-MSAs。PSUs 進行抽樣後再依據NAEP地區（東北部、東南部、中部、西部）進行分類，而每個地區包含1/4的人口，如圖4-1-2所示。

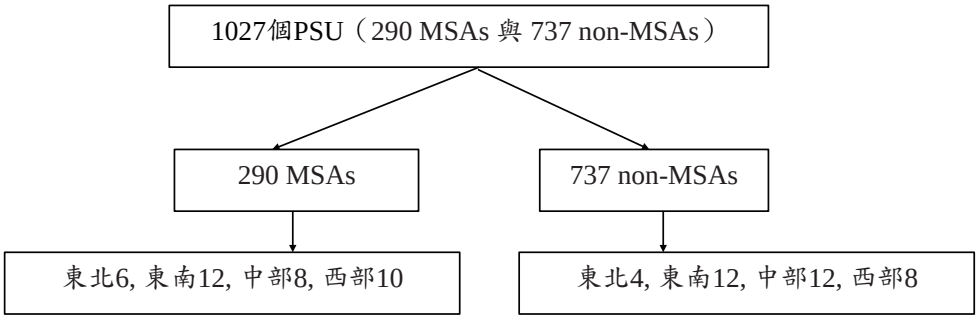


圖4-1-2　非必定抽樣之PSU分類

實際抽測94個PSUs：22個為必定會被抽取的PSUs（這些PSUs有較多的受試樣本，使得學校與學生能提供近似最佳解與符合成本效益的功能），而剩下非必定抽取之72個PSUs，則預計抽取36個MSAs與36個non-MSAs的PSUs。其中，36個MSAs與36個non-MSAs的PSUs皆依據該層級內的PSU個數，且透過按規模大小成比例機率的抽樣方法（probabilities proportional to size, PPS）進行抽樣，如表4-1-3所示。

表4-1-3　NAEP地區PSU分類

Region	Number of Strata for MSA PSUs	Number of Strata for Non-MSA PSUs	Total
Northeast	6	4	10
Southeast	12	12	24
Central	8	12	20
West	10	8	18
Total	**36**	**36**	**72**

資料來源：NAEP 1998 Technical Report, p.37

第二階段（抽取學校樣本）

施測學校是由每個PSU內的學校資料進行挑選，學校資料是透過列表架構與領域架構兩個來源所取得，其中，列表架構是根據幾個協會、州等提供的一些學校列表；領域架構則是補足不在列表架構之中的學校（上述這些學校名冊所指的是私立學校部分）。此外，公立、BIA、以及DoDEA等學校列表，則是由量化教育資料（Quality Education Data, QED）所提供。

各年級抽測學校透過PPS進行樣本學校之抽取，學校分層的變項主要包含：NAEP地區、公私立（private/public）分類、所在地（type of location）、高低少數民族（high/low minority）分類、PSU層級、以及施測學生年級等。分類順序是依據公私立學校、必然或非必然的PSU。其中，高少數民族指的是黑人與西班牙人，且在美國中小學，殘障學生和英文能力不佳的學生約占學生總數的10%，這兩類學生也會被選入NAEP的受試樣本中，各年級抽取學校數如表4-1-4所示。

表4-1-4　各年級抽取學校數

Grade	Region	MSA Certainty PSU	MSA Noncertainty PSU	Non-MSA Noncertainty PSU	Total
4	Northeast	125	54	17	196
	Southeast	27	105	61	193
	Central	78	80	59	217
	West	145	88	50	283
	Total	375	327	187	889
8	Northeast	142	60	18	220
	Southeast	29	110	70	209
	Central	90	84	62	236
	West	148	95	49	292
	Total	409	349	199	957
12	Northeast	122	45	19	186
	Southeast	29	101	79	209
	Central	68	59	55	182
	West	139	84	52	275
	Total	358	289	205	852

資料來源：NAEP 1998 Technical Report, p.41

其中，實際參與的學校數＝原始抽測學校數 - 超出範圍或倒閉的學校數 - 沒有適合受試學生的學校數 - 地區拒絕的學校數 - 學校拒絕的學校數 + 取代的學校。

表4-1-5　各年級實際參與的學校數

	Grade 4	Grade 8	Grade 12	Total	Public*	Nonpublic†
Total Original Sample	889	957	852	2,698	1,581	1,117
Out-of-Range or Closed	54	79	103	236	29	207
No Eligibles Enrolled	7	7	4	18	0	18
State Tested All Students	1	0	0	1	1	0
District Refused	52	50	50	152	151	1
School Refused	104	118	135	357	162	195
Cooperating	671	703	560	1,934	1,238	696
Cooperation Rate Before Substitution‡	81%	81%	75%	79%	80%	78%
(1996)	86%	83%	79%	83%	85%	80%
(1994)	86%	86%	79%	83%	82%	85%
(1992)	86%	85%	81%	84%	86%	82%
Cooperating Replacement for Refusals	62	58	48	168	109	59
Total Cooperating Schools	**733**	**761**	**608**	**2,102**	**1,347**	**755**
Cooperation Rate After Substitution	89%	87%	82%	86%	87%	85%
Total Students Assessed	**36,104**	**48,797**	**48,588**	**133,489**	**110,825**	**22,664**

資料來源：NAEP 1998 Technical Report, p.43

表4-1-6　NAEP學校類別

	Grade 4	Grade 8	Grade 12
Region			
Northeast	161	170	123
Southeast	174	175	167
Central	173	187	121
West	225	229	197
School Type			
Public	473	427	446
Private	93	114	82
Catholic	28	33	19
BIA	138	186	59
DODEA	1	0	2
Size and Type of Community			
Rural	157	166	113
Disadvantaged Urban	148	141	108
Advantaged Urban	192	209	153
Big City	49	54	45
Fringe	9	10	8
Medium City	80	76	77
Small Place	98	105	104
Number of Enrolled Students			
10-250	194	192	101
251-500	245	194	105
501-1000	208	209	106
1,001-2,000	28	91	158
2,000+	1	7	78

資料來源：NAEP 1998 Technical Report, p.444

第三階段（考科與樣本類型的分配）

考科類型：NAEP 1998考科類型包含writing/civics, civics special trend, and reading，其中，writing/civics考科4年級施測25分鐘的寫作與25分鐘的公民；8與12年級施測25分鐘的寫作、50分鐘的寫作、以及50分鐘的公民。因此，4年級包含4個考科類型、8與12年級包含5個考科類型。

樣本類型：NAEP除了一般受試者外，樣本類型還分為S2與S3。其中，S2指對於SD與LEP學生沒有提供特別的協助；S3指對於SD與LEP學生有提供特別的協助（例如：延長考試時間等）。

考科分配：以4年級為例，每個施測學校依據符合的施測學生數（如表4-1-7），分配一些考科，而考科的分配是重複以下的次序：R, W, W, W, R, W, W, W, R, W, W, W, R, W, W, C, W, W。其中，W為寫作/公民、R為閱讀、C為公民特別趨勢。

表4-1-7　施測學校考科數分配

Estimated Number of Grade-Eligible Students	Number of Sessions Allocated
1 – 25	1
26 – 50	2
51 – 75	3
76 or More	4

資料來源：NAEP 1998 Technical Report, p.44

舉例來說：若學校施測樣本數是78人，由表4-1-7可知該校應分配3個施測考科數。因此，學生數高於26人，則一定能分配施測到寫作/公民考科（W）；學生數高於76人，則幾乎能分配施測到閱讀考科（R）。此外，這個次序包含13個W、4個R與1個C，依據此方式能擔保考科W分配72%的施測學生數、考科R分配22%的施測學生數、考科C分配6%的施測學生數。

第四階段（受試學生選取）

NAEP 1998 每一間抽測學校最多抽取之受試者為：4年級100名、8年級125名、12年級150名。受試學生之抽取是根據抽樣學校準備的學生名冊進行抽樣，學生透過系統抽樣的方法，分配施測之考科類型。若學校學生數大於欲抽樣之樣本數，則隨機抽取各年級欲抽樣最多之樣本數；反之，學校學生數小於欲抽樣之樣本數，則全部學生皆參與施測。

二、PISA（OECD, 2005; OECD, 2006）

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。主要抽樣步驟為：定義各國的母群體、建立抽樣架構、確認各抽樣層級、學校樣本的分配、學校樣本的挑選（如圖4-1-3所示）。

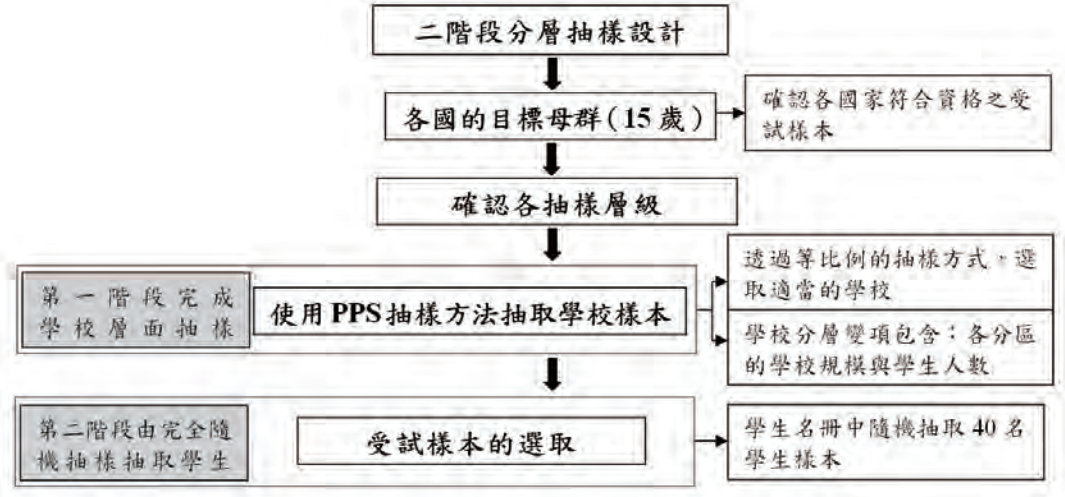


圖4-1-3　PISA抽樣架構

本書以PISA 2006技術報告為例，PISA目標母群為在所有參與施測國家中十五歲的學生（大部分學生的分布為九年級或是更高年級的學生）。舉例來說，PISA 2006對台灣進行施測，依據教育部統計處的資料顯示，臺灣地區符合PISA 2006 受測資格的全國滿十五歲的人口數為三十三萬四千三百九十一人，排除豁免測試的學生（如華僑學校、美國學校、啟智學校等），以及沒有在學（就業或補習）的學生，再經由PISA 的ACER與Westat二個組織確認，審查無誤後臺灣評量中心即依符合資格之學校分區及分學制列出符合資格之學生。PISA抽樣設計在大部分的國家為二階段的分層抽樣，第一階段完成學校層面的抽樣，總體來說，該階段是使用PPS的抽樣方法；第二階段是以完全隨機抽樣為原則。此外，有兩個國家使用三階段的抽樣設計，以地理區域為第一階段的抽樣，使用PPS抽樣方法；第二階段為抽樣學校，則依據所選取的地理區域抽取；第三階段則以學生為抽樣單位。接著，分別介紹PISA各階段抽樣方式。

第一階段（受測學校的選取）

評量中心除了必須備齊各校十五歲的受測學生數之外，也需同時呈報PISA各校相關資料，例如學校代號、網址、佔地面積、全校人數、電話等。抽樣專責單位嚴審確認資料無誤後，再配合各分區學校之學校規模、學生人數，分層抽樣代表受測的學校。其中，學校依據學校名冊並透過PPS的抽樣方式抽取出適當數量的學校。

第二階段（受測學生的選取）

首先，進行各校符合受測資格之學生名單收集，在確認各校所回傳之學生資料無誤後，即將相關資料匯入Keyquest 資料管理系統進行取樣。其中，此系統需輸入以下六項資料：（1）國家代碼、（2）施測語言、（3）就讀類科、（4）職業代碼、（5）學校資料、（6）符合受測學生資料。以上六項資料鍵入後，資料庫將作自動確認，將欄位重設整理，並選抽出每校需受測的四十名學生，這些學生來自不同的班級。其中，學生是由已確認的學生名冊中，以隨機抽樣方式選出大約 40位學生進行施測，並於施測3天前發放學生追蹤表（the student tracking form, STF）給參與施測之學生。

三、TIMSS（Mullis, Martin, Ruddock, O`Sullivan, Arora, & Erberber, 2005）

在國際教育成就調查委員會（International Association for the Evaluation of Educational Achievement, IEA）主辦的國際測驗中，TIMSS目標母群是指各國提供的母群體。TIMSS 2003樣本是由兩個目標母群中挑選，各國可以自由參加其中一個群體，或者是兩個都參加。目標母群包含四年級在籍的學生（大部分年齡為9歲）與八年級在籍的學生（大部分年齡為13歲）。TIMSS使用二階段分層之集群抽樣設計（two stage stratified cluster sample design），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。因此，TIMSS抽樣權重的計算根據三個程序選擇不同的機率，這三個程序包括學校、班級、以及班級內的抽樣（如圖4-1-4）。

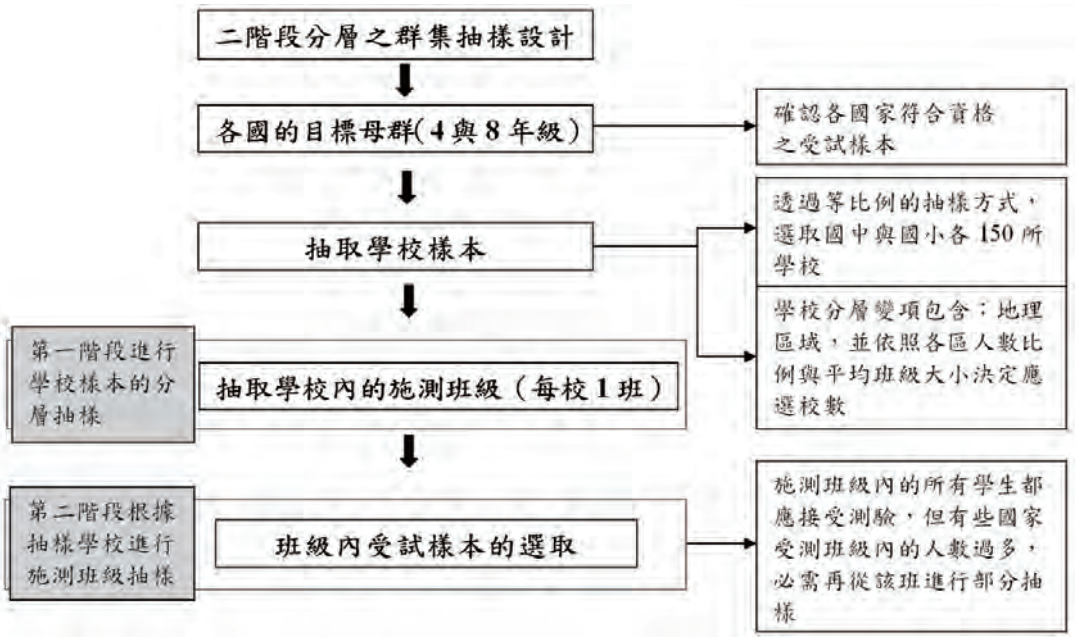


圖4-1-4　TIMSS抽樣架構

TIMSS 2003使用二階段分層之集群抽樣設計，第一階段是使用PPS的原則抽取國中與國小各150 所學校（4年級與8年級），使得能將學校層次與班級層次的平均估計之95%信賴區間控制在標準差的16%以內（張郁雯，2008）。舉例來說，TIMSS 2003在台灣施測時，首先將學校依據地理區域分成北、中、南、東四區，依照各區人數比例與各區平均班級大小決定各區應選校數，同時指定兩個遞補學校，以便當選取的學校拒絕加入研究時依序遞補。第二階段則由被抽到的學校中隨機抽取一個班級（Martin, Mullis, & Chrostowski, 2004）。

TIMSS 2003總計4年級抽取4661人，8年級抽取5379人。TIMSS考慮此抽樣特性，分析時使用TIMSS 2003資料庫中所提供的樣本加權變項中的house weight對樣本進行加權計算。依照TIMSS 2003 使用手冊的建議，因為每位學生的抽樣權值已知，進行分析時必須加權才能估計得到正確的母群估計值，使用house weight的優點為確保適當加權，但又將樣本數固定在實際抽樣人數，避免因為加權使得人數激增，影響統計顯著性考驗之正確性（Martin, 2005；張郁雯，2008）。接著，分別介紹TIMSS各階段抽樣方式。

TIMSS第一階抽樣單位是學校，第二階抽樣單位是受測學校內的班級，第三階抽樣單位是受測班級內的學生，基本上被抽到的受測班級內的所有學生都應接受測驗，但是有些國家的受測班級內的人數過多，必需再從該班進行部分抽樣，以亂數隨機選取適當的人數參加測驗。所以本階段的抽樣並不是每一個國家都需要使用。

此外，為了抽樣的精確度考量，各年級抽樣人數，原則上每個參與國個年級至少需有150間學校及至少4500位學生參加施測，若不達此要求的國家，在國際報告中會分開處理（譚克平，2009）。

四、TASA

TASA 2005對國民小學六年級學生，統一進行國語、英語、數學三科之成就評量，TASA 2007起施測對象增加為小四、小六、國二、高中二、高職二之學生，科目涵蓋國語文、英語文、數學、自然、社會領域（不含小四社會、2009起不含小四英語）五科。此外，TASA施測樣本不包含特殊學生，特殊學生則包括在家教育與12類障礙類別，其中，12類障礙類別包括：智能障礙、視覺障礙、聽覺障礙、語言障礙、肢體障礙、身體病弱、嚴重情障、學習障礙、多重障礙、自閉症、發展遲緩及其他顯著障礙。且因臺灣各縣市人口多寡各異，為充分顯現教改後臺灣學生學習成就實際情形，且確保TASA所抽取之樣本具有全國代表性，因此，採二階段隨機抽樣設計（國中小部份）。第一階段採分層叢集隨機抽樣，根據縣市、人口密度、班級數等三個變項進行分層。第二階段則根據所抽取到之樣本學校，對每一層以學生個人為單位，進行簡單隨機抽樣。在高中職部分則採全國學校普測，學生部份進行抽測，以比例進行運算（如圖4-1-5）。

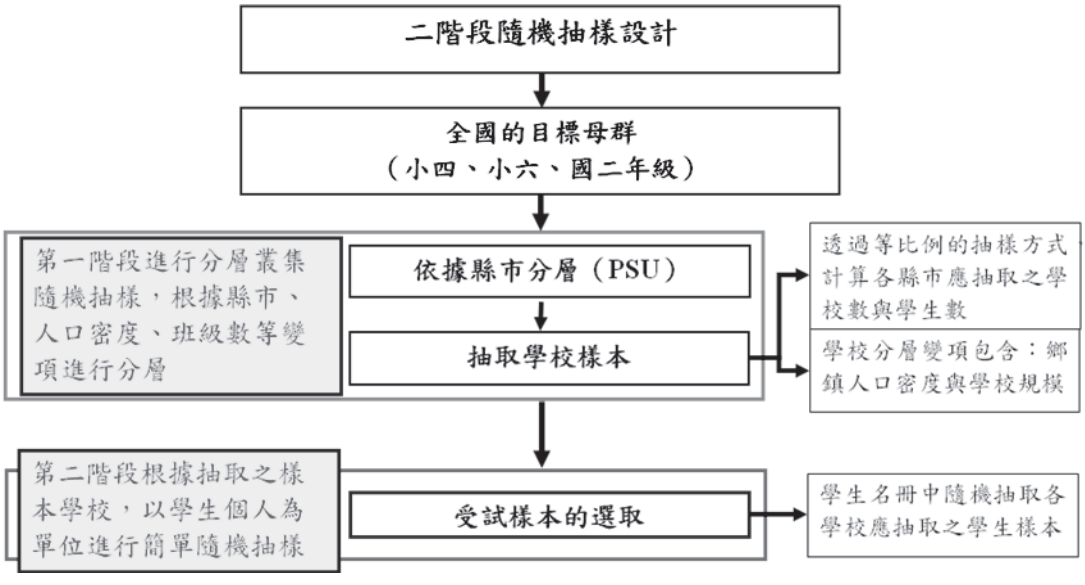


圖4-1-5　TASA抽樣架構

TASA分層設計主要分為二個步驟，分別說明如下：

第一階段（分層隨機抽樣，計算應抽取樣本學校數與學生數）

計算出各縣市所應抽取之樣本學校與人數，分別說明如下：

抽樣學校數：確認預計抽樣學校數。

抽樣學生數：根據各縣市各年級學生占全國當年級學生總數之百分比，計算各縣市應抽樣之學生數，此外，若每科不足200名受試者之縣市需補足縣市樣本。

第二階段（簡單隨機抽樣，進行學校分層設計）

TASA分層隨機抽樣，根據縣市、人口密度、學校規模三個變項進行分層，其中縣市依照行政區分為25個縣市；人口密度則以全縣人口平均數為基準，將鄉鎮（市、區）依照人口密度分為4群；至於學校規模則以24班以下及25班以上為基準，分為2群，因此，整個抽樣成為一個25×4×2之分層設計。

五、綜合討論與建議

綜合上述，可知大型測驗主要皆是透過多階段的抽樣方法抽取受試樣本，主要分成兩個階段，包含受試學校與受試學生的選取。因此，本書主要目的是藉由探討大型測驗的抽樣架構，以檢視TASA的抽樣設計。表4-1-8整理上述大型測驗之抽樣設計，由於TASA抽樣設計的具體建議內容較多，因此，將整理成第伍部分統一說明。

表4-1-8　大型測驗之抽樣設計

NAEP	PISA	TIMSS	TASA
多階段分層抽樣設計，主要分為四個階段：第一階段抽樣單位是郡、第二階段抽樣單位是學校、第三階段抽樣為學校考科類型與樣本類型的分配、第四階段為學生選取與考科類型的分配。	二階段分層抽樣的抽樣設計：第一階段完成學校層面的抽樣；第二階段為完全隨機抽樣。	二階段分層之集群抽樣設計：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。	二階段隨機抽樣設計(國中小部份)，第一階段進行分層叢集隨機抽樣；第二階段根據抽取之樣本學校，以學生個人為單位，進行簡單隨機抽樣。

貳、抽樣權重

一、抽樣權重之計算

以下透過簡單的例子，介紹抽樣權重之計算。舉例來說，有10間學校，每間學校有40位學生（共400位學生），假設將抽取4間學校，且被抽取之學校將選10位受試者。因此，學校i被選取的機率為：$P_{1\_i}=\frac{n_{sc}}{N_{sc}}=0.4$，其中，$n_{sc}$為預計抽取之學校數，$N_{sc}$為所有的學校數；被抽取之學校內，學生$j$被選取的機率為：$P_{2\_ij}=\frac{n_i}{N_i}=0.25$，其中，$N_i$為學校$i$中之學生數；$n_i$為學校$i$中之抽樣學生數。這指在每個被抽取之樣本學校中，每位學生有0.25的機率被選取；學生j於學校i中的最終機率為：$P_{ij}=P_{1\_i}P_{2\_ij}$，在這個例子中，$P_{ij}=0.1$。接著，定義w_{1_i}為學校權重；w_{2_ij}為學校內權重；以及w_{ij}為學校最終的權重總合，算式如下：$w_{1\_i}=\frac{1}{P_{1\_i}}=\frac{1}{0.4}=2.5$；$w_{2\_ij}=\frac{1}{P_{2\_ij}}=\frac{1}{0.25}=4$；$w_{ij}=\frac{1}{P_{ij}}=\frac{1}{0.1}=10$。由表4-1-9可知，學校權重的總和會等於學校的總數；最後權重的總和會等於母群體的人數。然而，在實際上每間學校的人數並不相同，城市地區的學生人數相對於農村地區多。若學校的選取採用簡單隨機抽樣，則學校被選取的機率將沒有改變。但在被選取的學校內，學生被選取的機率則取決於學校人數的大小（表4-1-10）。

表4-1-9　學校、學校內、最後機率與相對應之權重（各學校的學生數相同）

School label	School size N_i	School prob. p_{1_i}	School weight w_{1_i}	Within-school prob. P_{2_ij}	Within-school weight w_{2_ij}	Final student prob. p_{ij}	Final student weight w_{ij}	Sum of final weights $n_i w_{ij}$
1	40							
2	40	0.4	2.5	0.25	4	0.1	10	100
3	40							
4	40							
5	40	0.4	2.5	0.25	4	0.1	10	100
6	40							
7	40	0.4	2.5	0.25	4	0.1	10	100
8	40							
9	40							
10	40	0.4	2.5	0.25	4	0.1	10	100
Total			10					400

資料來源：PISA 2003 Data Analysis Manual, p.23

表4-1-10　學校、學校內、最後機率與相對應之權重（各學校的學生數不同）

School label	School size	School prob.	School weight	Within-school prob.	Within-school weight	Final student prob.	Final student weight	Sum of final weights
1	10							
2	15	0.4	2.5	0.66	1.5	0.27	3.75	37.5
3	20							
4	25							
5	30	0.4	2.5	0.33	3	0.13	7.5	75
6	35							
7	40	0.4	2.5	0.25	4	0.1	10	100
8	45							
9	80							
10	100	0.4	2.5	0.1	10	0.04	25	250
Total	**400**		**10**					**462.5**

資料來源：PISA 2003 Data Analysis Manual, p.24

若使用簡單隨機抽樣於不同人數大小的學校中，所有學校將有相同的被選取機率，因此，學校權重總和會等於學校總個數。然而，學生最終的權重總和不一定等於母群體的人數，且學生最終的權重在不同的學校之中也不相同。這個變異將減少所有母群體參數估計的信度。表4-1-11與表4-1-12呈現不同的情形，由此可知學生最終的權重總和與期望值400有相當大的差異，而學校權重總和仍等於學校總個數。

表4-1-11　學校、學校內、最後機率與相對應之權重（選擇小樣本之學校）

School label	School size	School prob.	School weight	Within-school prob.	Within-school weight	Final student prob.	Final student weight	Sum of final weight
1	10	0.4	2.5	1	1	0.4	4	40
2	15	0.4	2.5	0.66	1.5	0.27	3.75	37.5
3	20	0.4	2.5	0.5	2	0.2	5	50
4	25	0.4	2.5	0.4	2.5	0.16	6.25	62.5
Total			**10**					**190**

資料來源：PISA 2003 Data Analysis Manual, p.25

表4-1-12　學校、學校內、最後機率與相對應之權重（選擇大樣本之學校）

School label	School size	School prob.	School weight	Within-school prob.	Within-school weight	Final student prob.	Final student weight	Sum of final weight
7	40	0.4	2.5	0.250	4	0.10	10.00	100.0
8	45	0.4	2.5	0.222	4.5	0.88	11.25	112.5
9	80	0.4	2.5	0.125	8	0.05	20.00	200.0
10	100	0.4	2.5	0.100	10	0.04	25.00	250.0
Total			10					662.5

資料來源：PISA 2003 Data Analysis Manual, p.25

許多國際的教育測量重視受試樣本多於學校樣本；然而，由簡單隨機抽樣進行學校樣本選取之抽樣設計是不適當的，因為它將低估或高估母群體的個數，而增加抽樣變異（sampling variability）。PISA 為了避免這些缺點，學校的選取是透過 PPS 的抽樣方法，受試者人數多的學校相對於受試者人數少的學校有較高的被選取機率，但學生於受試者人數多的學校相對於受試者人數少的學校有較低的被選取機率。

因此，依據式子 $P_{1\_i} = \frac{N_i \times n_{sc}}{N}$ ，

可計算學校9被選取的機率為 $P_{1\_9} = \frac{N_9 \times n_{sc}}{N} = \frac{80 \times 4}{400} = 0.8$ ；

學生於學校9被選取的機率為 $P_{2\_9j} = \frac{n_9}{N_9} = \frac{10}{80} = 0.125$ ；

學生最終權重為 $P_{9j} = 0.8 \times 0.125 = 0.1$ 。

由表4-1-13可知這樣的設計不會增加抽樣變異，且最後權重的總和會等於母群體的人數，但學校權重總和並沒有等於學校總個數（若選取學校人數最多的4間學校，學校權重總和為6.97；若選取學校人數最少的4間學校，學校權重總和為25.67）；然而，這並不會是教育測量的主要問題，主要感興趣的仍是學生樣本。為了使這個差異能達到最小（學校個數與學校權重總和），學校的選取將透過系統化的程序。

系統化程序首先將學校依人數多少排序，抽樣區間是依據母群體人數與抽樣學校數的比值，$Int = \frac{N}{n_{sc}} = \frac{400}{4} = 100$ 。由均勻分布[0, 1]中隨機選取一個數字，假設選取為0.752，將此數字乘上抽樣區間（0.752×100=75.2）。則第75.2位受試者所在的學校將被選取，且接下來選取之學校是第75.2位受試者後每間隔100位受試者其所在的學校將被選取，如表4-1-14所示。

表4-1-13　學校、學校內、最後機率與相對應之權重於PPS抽樣方法

School label	School size	School prob.	School weight	Within-school prob.	Within-school weight	Final student prob.	Final student weight	Sum of final weight
1	10							
2	15							
3	20	0.2	5.00	0.500	2.0	0.1	10	100
4	25							
5	30							
6	35							
7	40	0.4	2.50	0.250	4.0	0.1	10	100
8	45							
9	80	0.8	1.25	0.125	8.0	0.1	10	100
10	100	1	1.00	0.100	10.0	0.1	10	100
Total	400		9.75					400

資料來源：PISA 2003 Data Analysis Manual, p.26

表4-1-14　以PPS抽樣進行樣本學校之選擇

School label	School size	From student number	To student number	Part of the sample
1	10	1	10	No
2	15	11	25	No
3	20	26	45	No
4	25	46	70	No
5	30	71	100	Yes
6	35	101	135	No
7	40	136	175	No
8	45	176	220	Yes
9	80	221	300	Yes
10	100	301	400	Yes

資料來源：PISA 2003 Data Analysis Manual, p.27

二、綜合討論與建議

由上述可知若使用兩階段抽樣設計，即可以擔保所有受試樣本有相同被選取的機率，也就代表所有受試樣本有相同的抽樣權重。然而，若是如此為何大型測驗仍然需要計算抽樣權重呢？PISA 2003資料分析手冊中指出（OECD，2005），由於實際施測時有超取樣或取樣不足、抽測樣本與實際樣本不同、以及學校與學生無作答反應權重調整等的情況發生，使得大型測驗仍然需要計算抽樣權重。因此，有關TASA抽樣設計權重之計算，以及建議如何調整TASA抽設權重，也將整理成第伍部分統一說明。

參、抽樣變異估計方法

全國性或國際性的測量在進行收集資料時，通常會使用抽樣來代替普查。某些特定的母群可能包含幾千個甚至上百萬個樣本，且他們不一定有相同的母群統計估計值。且每一個母群統計的估計值會有一個關聯的不確定性或是誤差風險，而抽樣變異相當於測量這個不確定性。因此，此部分將探討如何使用複製權重（replicate weights）於一個複雜的抽樣設計中，以估計母群的抽樣變異。

一般而言，兩階段抽樣設計有三種複製（ replication）方法類別來估計抽樣變異，包含：Jackknife方法（分為有階層的抽樣與沒有階層的抽樣）、平衡重複方法（balanced repeated replication, BRR）、以及Bootstrap方法。由於TIMSS與NEAP使用Jackknife方法估計抽樣變異，而PISA使用修正BRR的方法估計抽樣變異，因此，以下僅介紹這幾種抽樣變異估計方法。

一、Jackknife方法

若假設抽取PSUs使用簡單隨機抽樣且沒有使用任何層級變數的情況下，則使用Jackknife方法獲得的平均抽樣變異會等同於使用簡單隨機抽樣進行兩階段抽樣設計的抽樣變異，如下式：

$$\sigma^2_{(\hat{\mu})} = \frac{\sigma^2_{between\_PSU}}{n_{PSU}} + \frac{\sigma^2_{within\_PSU}}{n_{PSU}\ n_{within}}$$

因此，假設有10間抽樣學校，且每個學校內的學生是由簡單隨機抽樣所獲得，則Jackknife方法在這一個沒有層級的兩階段抽樣設計中，由10個replicates產生9間學校。在系統方法中，每個學校僅去除一次。

表4-1-15　Jackknife方法使用於無階層之兩階段抽樣設計

Replicate	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
School 1	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
School 2	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11
School 3	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11	1.11
School 4	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11	1.11
School 5	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11	1.11
School 6	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11	1.11
School 7	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11	1.11
School 8	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11	1.11
School 9	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00	1.11
School 10	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	1.11	0.00

資料來源：PISA 2003 Data Analysis Manual, p.47

在第一個複製（R1）中，學校1是被移除，且其他學校的權重被調整為1.1（$\frac{G}{G-1}=\frac{10}{9}$）。這個調整因素被應用在，當學校複製權重與學校內的複製權重被結合成學生複製權重時；在第二個複製中，學校2是被移除，且其他的學校被調整成相同的權重…依此類推。因此，replicate估計值的抽樣變異，如下：

$$\sigma^2_{(\hat{\theta})} = \frac{G-1}{G}\sum_{i=1}^{G}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

然而，在有層級的兩階段抽樣設計下，Jackknife的方法會使抽樣變異減少，且將導致一個系統高估的抽樣變異。因此，必須透過PPS的抽樣方法與系統程序進行受試學校的挑選，以及定義層級變數等實施步驟。舉例來說，假設母群的列表學校分成兩個部份：農村學校與城市學校，且在這兩個層級學校依據學生大小排序。在每個層級內，10間學校透過系統程序與比例的抽樣方法進行挑選。因此，在有層級的兩階段抽樣設計下，Jackknife的方法是在每個層級內依照學校被挑選的次序，進行有系統地配對樣本學校。因此，學校將與其他相似的學校配對。

表4-1-16　Jackknife方法使用於階層之兩階段抽樣設計

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R10
1	1	2	1	1	1	1	1	1	1	1	1
1	2	0	1	1	1	1	1	1	1	1	1
2	3	1	0	1	1	1	1	1	1	1	1
2	4	1	2	1	1	1	1	1	1	1	1
3	5	1	1	2	1	1	1	1	1	1	1
3	6	1	1	0	1	1	1	1	1	1	1
4	7	1	1	1	0	1	1	1	1	1	1
4	8	1	1	1	2	1	1	1	1	1	1
5	9	1	1	1	1	2	1	1	1	1	1
5	10	1	1	1	1	0	1	1	1	1	1
6	11	1	1	1	1	1	2	1	1	1	1
6	12	1	1	1	1	1	0	1	1	1	1
7	13	1	1	1	1	1	1	0	1	1	1
7	14	1	1	1	1	1	1	2	1	1	1
8	15	1	1	1	1	1	1	1	0	1	1
8	16	1	1	1	1	1	1	1	2	1	1
9	17	1	1	1	1	1	1	1	1	0	1
9	18	1	1	1	1	1	1	1	1	2	1
10	19	1	1	1	1	1	1	1	1	1	2
10	20	1	1	1	1	1	1	1	1	1	0

資料來源：PISA 2003 Data Analysis Manual, p.48

表4-1-16呈現如何產生複製，學校1~10為農村學校，學校11~20為城市學校。在每個層級裡都有5個學校配對。Jackknife的方法產生許多replicates，在這個例子中，產生10個replicates，對於每個replicate樣本，隨機選擇一間學校將它移除，並給剩餘的那間學校兩倍的權重。在第一個複製（R1）中，學校2是被移除且學校1有2倍的權重（pseudo-stratum 1）；在第二個複製中，學校3是被移除且學校4有2倍的權重（pseudo-stratum 2），依此類推。而抽樣變異，如下式：

$$\sigma^2_{(\hat{\theta})} = \sum_{i=1}^{G}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

二、BRR方法

Jackknife方法在每一個複製的樣本中僅移除一個學校，BRR的方法是在每一個虛擬層級（pseudo-stratum）內隨機選擇一間學校，並將它的權重設為0，且針對剩餘的學校給定加倍的權重。此方法導致一大堆可能的複製，平衡的複製樣本產生是依據阿達馬矩陣（Hadamard matrices），這是為了避免冗長的計算，而複製的次數是大於或等於pseudo-strata數量的最小4的倍數。在這個例子中，共有10個pseudo-strata，因此，需產生12個複製。則抽樣變異，如下式：

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G}\sum_{i=1}^{G}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

表4-1-17　BRR複製方法

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12
1	1	2	0	0	2	0	0	0	2	2	2	0	2
1	2	0	2	2	0	2	2	2	0	0	0	2	0
2	3	2	2	0	0	2	0	0	0	2	2	2	0
2	4	0	0	2	2	0	2	2	2	0	0	0	2
3	5	2	0	2	0	0	2	0	0	0	2	2	2
3	6	0	2	0	2	2	0	2	2	2	0	0	0
4	7	2	2	0	2	0	0	2	0	0	0	2	2
4	8	0	0	2	0	2	2	0	2	2	2	0	0
5	9	2	2	2	0	2	0	0	2	0	0	0	2
5	10	0	0	0	2	0	2	2	0	2	2	2	0
6	11	2	2	2	2	0	2	0	0	2	0	0	0
6	12	0	0	0	0	2	0	2	2	0	2	2	2
7	13	2	0	2	2	2	0	2	0	0	2	0	0
7	14	0	2	0	0	0	2	0	2	2	0	2	2
8	15	2	0	0	2	2	2	0	2	0	0	2	0
8	16	0	2	2	0	0	0	2	0	2	2	0	2
9	17	2	0	0	0	2	2	2	0	2	0	0	2
9	18	0	2	2	2	0	0	0	2	0	2	2	0
10	19	2	2	0	0	0	2	2	2	0	2	0	0
10	20	0	0	2	2	2	0	0	0	2	0	2	2

資料來源：PISA 2003 Data Analysis Manual, p.49

這個複製的方法，每個複製樣本只使用一半有效的觀察值。然而，這樣大量的減少樣本可能會造成極端子群統計估計的問題。此外，一些剩餘的觀察值可能是比較小的（甚至等於0），則對於特定複製樣本母群參數估計是不可能的。

為了克服這個不利的條件，Fay 發展一個不同的BRR方法，代替原本的權重0與2。Fay 建議權重依據一個緊縮的係數k（介於0~1），而第二個膨脹的係數則為2減去k。例如：假設緊縮係數k為0.6，則膨脹權重係數為1.4（Judkins, 1990）。PISA使用Fay 的方法，且將k設為0.5，如表4-1-18所示。則抽樣變異，如下式：

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}\left(\hat{\theta}_{(i)} - \hat{\theta}\right)^2$$

表4-1-18　Fay複製方法

Pseudo-stratum	School	R1	R2	R3	R4	R5	R6	R7	R8	R9	R 10	R 11	R 12
1	1	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5
1	2	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5
2	3	1.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5
2	4	0.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5
3	5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5
3	6	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5
4	7	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5	1.5
4	8	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5	0.5
5	9	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5	1.5
5	10	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5	0.5
6	11	1.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5	0.5
6	12	0.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5	1.5
7	13	1.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5	0.5
7	14	0.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5	1.5
8	15	1.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5	0.5
8	16	0.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5	1.5
9	17	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5	1.5
9	18	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5	0.5
10	19	1.5	1.5	0.5	0.5	0.5	1.5	1.5	1.5	0.5	1.5	0.5	0.5
10	20	0.5	0.5	1.5	1.5	1.5	0.5	0.5	0.5	1.5	0.5	1.5	1.5

資料來源：PISA 2003 Data Analysis Manual, p.50

在PISA中，決定產生80個複製樣本與80個複製權重，因此，計算如下：

$$\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-k)^2}\sum_{i=1}^{G}\left(\hat{\theta}_{(i)}-\hat{\theta}\right)^2 = \frac{1}{80(1-0.5)^2}\sum_{i=1}^{80}\left(\hat{\theta}_{(i)}-\hat{\theta}\right)^2$$

$$= \frac{1}{20}\sum_{i=1}^{80}\left(\hat{\theta}_{(i)}-\hat{\theta}\right)^2$$

三、綜合討論與建議

教育研究與許多特別國際性研究學生樣本的抽取，通常使用兩階段抽樣設計，先選取學校的樣本，再由每個選取的學校中隨機選取班級或學生。而在相同學校中學生被選取不能視為獨立的觀測值，因為學生在相同的學校通常比學生在不同的學校有較多的共同特徵，例如：他們提供相同的學校資源，可能是相同的老師，以及教授相同的課程等。因此，不同學校中學生的差異有可能是比較大的，例如：可以預期抽取一個職業學校（vocational school）的學生樣本與一個一般學校（academic school）的學生樣本會比抽取兩個一般學校的學生樣本有較多的變異。因此，兩階段抽樣設計的抽樣變異與第一階段抽樣單位（PSU）的差異呈直接比例。所以TIMSS與NAEP皆使用有階層抽樣的Jackknife方法估計抽樣變異，PISA則使用Fay的方法估計抽樣變異，由於Fay的方法能減少樣本可能會造成極端子群統計估計的問題，因此，本書亦建議使用Fay的方法估計TASA的抽樣變異。

肆、現行TASA抽樣設計的缺點

TASA使用二階段分層隨機抽樣設計，根據縣市、人口密度、學校規模三個變項進行分層，其中，縣市依照行政區分為25個縣市；人口密度以全縣人口平均數為基準，將鄉鎮（市、區）依照人口密度分為4群；學校規模則以24班以下及25班以上為基準，分為2群。本書藉由國外大型測驗（NAEP、PISA、TIMSS）抽樣設計之探討，認為TASA的抽樣設計將產生以下幾個問題：

1. TASA進行鄉鎮人口密度之分層，人口密度等級依據縣市各自的人口密度與縣市內各鄉鎮市人口密度進行分類，並畫分成四個等級。若以臺北縣與宜蘭縣為例：臺北縣總人口數為3736677人，總面積為2052.5667平方公里，因此，每平方公里之人口密度約為1820（3736677/2052.5667）。接著，計算臺北縣各鄉鎮市人口密度，臺北縣共計29個鄉鎮市，根據各鄉鎮市之人口數與面積數，可計算得各鄉鎮市之人口密度。其中，人口密度高於1820的鄉鎮市分為1與2層級，人口密度低於1820的鄉鎮市分為3與4層級；宜蘭縣總人口數為461586人，總面積為2143.63平方公里，因此，每平方公里之人口密度約為215，則人口密度高於215的鄉鎮市分為1與2層級，人口密度低於215的鄉鎮市分為3與4層級。因此，可知這樣的層級分類在各縣市間並無意義，例如：宜蘭縣層級1的鄉鎮市地區人口密度對應到的是臺北縣層級2與層級3的鄉鎮市地區。
2. 現行TASA抽樣設計使用非隨機的抽樣方法，依據預計抽取之學校數與學生數進行等比例的計算，分別計算各縣市應施測之學校數與學生數。然而，此抽樣方式學生被抽取之機率並不相同，不應該假設受試學生有相同之抽樣權重，且學校的選取未使用PPS的抽樣方式進行抽樣，使得學生最終權重值將受影響（有高估或低估的情況發生）。

伍、TASA抽樣設計建議方案

為因應現行TASA抽樣設計的缺點，共召開兩次諮詢會議探討TASA抽樣設計。與會者認為自2005年起施測至今，在抽樣學校行政作業上能克服以個人為抽樣單位，且由實徵資料分析若以個人為抽樣單位，學生最終權重能推論至母群體的學生；若以班級為抽樣單位，學生最終權重將受班級人數所影響（會有高估或低估的情況發生）。因此，在諮詢委員建議下，提出之TASA抽樣設計建議方案將以個人為抽樣單位進行抽樣。

本書將提出兩種TASA抽樣設計建議方案，在方案一中，建議TASA抽樣設計應刪除鄉鎮人口密度層級，並參考NAEP抽樣架構（Allen, Donoghue, & Schoeps, 2001），在抽取樣本學校後，先進行考科的分配，且樣本學校之選取是透過PPS的抽樣方法（詳見圖4-1-6）。此外，為了因應2010年12月25日起台灣行政規劃區域重新劃分，方案二建議將抽樣的PSU區分為5都（臺北市、新北市、臺中市、臺南市、高雄市）、4個地理區（北部、中部、南部、東部）、以及離島（澎湖、金門、馬祖），詳見圖4-1-7。

一、TASA 抽樣設計建議方案一

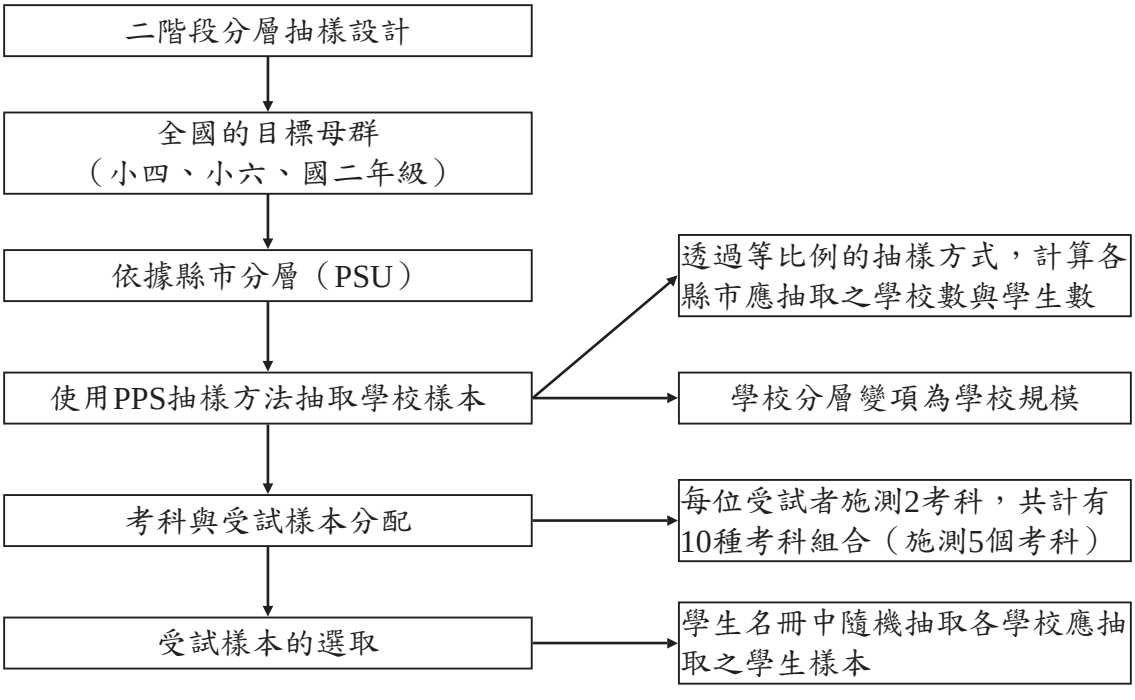


圖4-1-6　方案一的抽樣架構

本書建議TASA使用二階段分層抽樣設計（國中與國小的部分），第一階為分層叢集隨機抽樣，根據縣市與班級數兩個變項進行分層；第二階段再根據所抽取的樣本學校，以學生個人為抽樣單位進行簡單隨機抽樣。茲將本書建議之抽樣程序描述如下：

1. 選取樣本學校

TASA抽樣選取樣本學校分成兩階段，第一階段先抽取主要抽樣單位（PSU），第二階段再抽取學校樣本，分別說明如下（參考NAEP抽樣方式）：

(1) TASA定義25個縣市的行政區皆為必然抽取之PSU，然而，PSU內所需抽取之受試樣本是依據各縣市所占學生數之比例。

(2) 施測學校數是由每個PSU內所提供的學校資料進行挑選，各年級抽測學校數是透過規模大小所占學生數之比例進行抽樣（probabilities proportional to assigned measures of size），學校依據學校規模進行分層，以學校班級總數24班以下與25班以上為基準，分為2群。接著，透過按規模大小成比例機率的抽樣方法（probabilities proportional to size, PPS）進行施測學校之抽樣（參考PISA抽樣方式）。

2. 考科與受試樣本分配

TASA國小四年級受試者必須施測國語文、數學科、自然科等三科考科；國小六年級、國中二年級、高中二年級及高職二年級受試者必須施測國語文、英語文、數學科、自然科、社會科等五科考科，為考量受試者施測時間過長，造成疲勞等因素，每名受測學生自該年級應施測之考科中抽測二考科。因此，國小四年級包含3種考科組合，分別為：國數（V1）、國自（V2）、數自（V3）；其他年級包含10種考科組合，分別為：國英（T1）、國數（T2）、國自（T3）、國社（T4）、英數（T5）、英自（T6）、英社（T7）、數自（T8）、數社（T9）、自社（T10），為使各考科能有足夠且相同數量的受試者參與施測，抽樣架構擬加入考科分配。

考科分配：以國小六年級為例，每個施測學校依據符合的施測學生數（如表4-1-19），分配一些考科，而考科的分配是重複以下的次序：T1, T2, T3, T4, T5, T6, T7, T8, T9, T10。舉例來說：若學校施測樣本有48人，則此樣本學校將分配到2個考科組合數；若學校施測樣本有28人，則此樣本學校將分配到1個考科組合數。因此，若樣本學校1施測樣本有60人、樣本學校2施測樣本有53人、樣本學校3施測樣本有29人，則樣本學校1施測考科組合T1與T2、樣本學校2施測考科組合T3與T4、樣本學校3施測考科組合T5。

表4-1-19　施測學校考科數分配

樣本學校內被選取之施測學生數	分配考科組合之個數
1-30	1
31-60	2

3. 選取受試樣本

TASA 每一間抽測學校最多抽取之受試者為60名，受試學生依據抽樣學校提供之學生名冊進行抽樣，受試樣本是採用簡單隨機抽樣進行選取。若學校學生數大於欲抽樣之樣本數，則隨機抽取各年級欲抽樣最多之樣本數；反之，學校學生數小於欲抽樣之樣本數，則全部學生皆參與施測。

4. 抽樣程序（以TASA 2007 國中二年級為例）

TASA 預計各學科抽取7500位受試者，共抽取18750名受試者，依據教育部統計處提供的資料，2006年國中實際學校數為918間，其中，25班以上的學校有465間（50.65%），24班以下的學校有453間（49.35%），且班級數低於2班的學校共57間學校（6.21%）。因此，假設每間樣本學校隨機抽取60位受試樣本（2班），則須抽取313間樣本學校（18750 / 60 = 312.5）。接著依據班級數低於2班學校之比例，計算多抽取之學校數（313*0.0621=19.44）。是故，設定抽取學校數為350間。

由於推論至全國受試樣本之權重程序較為複雜，因此，目前預定進行新抽樣方式權重之計算，抽樣程序如下：

(1) 依據各縣市學生數所占之比例，計算各縣市應抽取之學校數。
(2) 依據各學校規模所占學生數之比例，計算應抽取之學校數。
(3) 透過PPS取樣方法抽取各縣市之樣本學校。
(4) 由樣本學校之學生名冊或班級名冊隨機抽取樣本學生或班級。

二、TASA 抽樣設計建議方案二（以五都為抽樣單位進行抽樣）

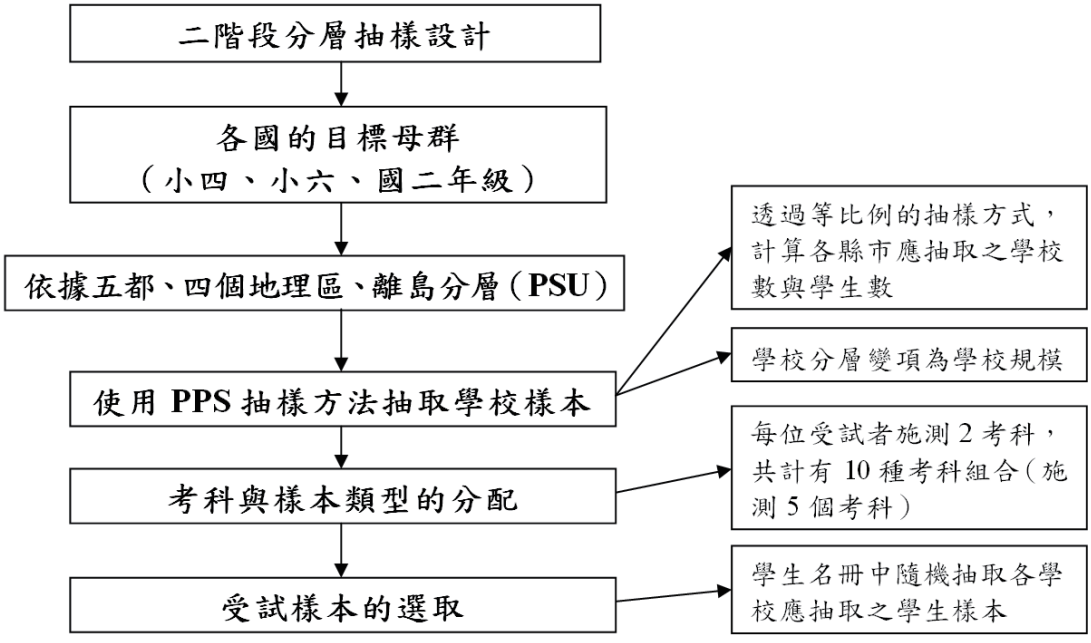


圖4-1-7 方案二的抽樣架構

由於2010年台灣行政地區將有所改變，且為了進行縣市或地區性受試樣本成就表現之比較，方案二依據五都、四個地理區、離島地區定義PSU，建議抽樣程序如下：

1. 選取樣本學校

TASA抽樣選取樣本學校分成兩階段，第一階段先抽取PSU，第二階段再抽取學校樣本，分別說明如下（參考NAEP抽樣方式）：

(1) TASA依據5都（臺北市、新北市、臺中市、臺南市、高雄市）、4個地理區（北部、中部、南部、東部）、以及離島（澎湖、金門、馬祖）定義為必然抽取之PSU（共10個PSU）。然而，PSU內所需抽取之受試樣本是依據所占學生數之比例。其中，北部：基隆市、桃園縣、新竹市、新竹縣；中部：苗栗縣、南投縣、彰化縣、雲林縣；南部：嘉義縣、嘉義市、屏東縣；東部：宜蘭縣、花蓮縣、臺東縣；離島：澎湖縣、金門縣、連江縣。

(2) 施測學校是由每個PSU內所提供的學校資料進行挑選，使用PPS的抽樣方法。

2. 選取受試樣本

TASA 每一間抽測學校最多抽取之受試者為60名，受試學生依據抽樣學校提供之學生名冊進行抽樣，受試樣本是採用簡單隨機抽樣進行選取。若學校學生數大於欲抽樣之樣本數，則隨機抽取各年級欲抽樣最多之樣本數；反之，學校學生數小於欲抽樣之樣本數，則全部學生皆參與施測。

3. 抽樣程序（以TASA 2007 國中二年級為例）

表4-1-20　計算各PSU應抽取之學校數

地區	學生數	百分比	學校數
臺北市	33347	10.43%	36
新北市	52483	16.42%	57
臺中市	41599	13.01%	46
臺南市	26459	8.28%	29
高雄市	36877	11.54%	40
北部地區	47459	14.85%	52
中部地區	41952	13.12%	46
南部地區	22434	7.02%	25
東部地區	15177	4.75%	17
離島地區	1879	0.59%	2
總計	319666	100.00%	350

第二節　測量模式

大型測驗的主要目的為評量群體的知識與技能，而評量牽涉到許多議題，首先進行測量前，應先編製所欲評量目標之試題，因應不同的評量目標需求，利用不同的試題類型進行施測，並確保試題的數量是夠多且可以涵蓋不同的難度範圍。施測時透過試題及題本的設計，受試者只施測部分的試題，以減輕受試者的負擔。

依據不同的試題類型，透過測量模式，提供不同的估計方法與模式的估計，將估計出來的試題與能力參數提供給受試者以及研究者。本節中，將介紹不同大型測驗中所使用的測量模式，針對模式估計的適合度進行探討，並羅列所使用估計試題與能力參數的測驗分析軟體。

以下將以NAEP 1998 （Allen, Carlson, Johnson, & Mislevy, 1999）、TIMSS 2007（Foy, Galia, & Li, 2008）和PISA 2003（OECD, 2005）的技術報告為主，針對這三大測驗所使用的試題類型、測量模式、模式適合度評估指標及測驗分析軟體作一整理說明。

壹、試題類型

目前在NAEP 1998（Allen, Carlson, Johnson, & Mislevy, 1999）、TIMSS 2007（Foy, Galia, & Li, 2008）、PISA 2003（OECD, 2005）和TASA大型測驗中，測驗的題型大致上可以分為三大類：

一、選擇題（multiple-choice items）

在四種大型測驗中皆為四個選項的選擇題。

二、填充題

NAEP的填充題（short constructed response items）可以分為答對答錯的二元計分，以及三點計分（0-2）兩種，PISA的填充題可以分為封閉性填充題（closed-constructed response items）以及開放性填充題兩種，封閉性填充題是指固定單一答案，像是數學科中，需要學生填入的可能為一個數值，開放性的填充題則是學生的反應可以較廣泛的作答，而非單一的答案，TIMSS的填充題（constructed-response items with just two response options）為二元計分的封閉性填充題。

三、開放性試題

NAEP中的開放性試題可以分為三點計分（0-2）到六點計分（0-5）四種，PISA中的開放性試題包括比較長的寫作、結論、摘要、批判等學生作答反應較廣泛之試題，TIMSS中的開放性試題，大多為三點計分（0-2），各技術報告中的名詞如表4-2-1所示：

表4-2-1　NAEP、PISA、TIMSS、TASA試題類型

NAEP	PISA	TIMSS	TASA
選擇題（multiple-choice items）	選擇題（multiple-choice response）	選擇題（multiple-choice items）	選擇題
填充題（short constructed response items）	填充題（short answer）	填充題（constructed response items with just two response options）	無此題型
開放性試題（extended constructed response items）	開放性試題（open constructed response）	開放性試題（constructed response items）	開放性試題

貳、測量模式

不同大型測驗間針對不同測驗題型，使用不同的測量模式，常見的有二參數對數模式（two-parameter logistic model, 2PL）、三參數對數模式（three-parameter logistic model, 3PL）、一般化部分給分模式（generalized partial credit model, GPCM）以及多向度隨機係數多項洛基模式（multidimensional random coefficients multinomial logit model, MRCMLM）。

一、二參數對數模式（2PL）

在IRT的2PL模式下，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）：

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

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），$-\infty<a_i<\infty$；b_i 為試題 i 之試題難度參數，$-\infty<b_i<\infty$。

二、三參數對數模式（3PL）

在IRT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）：

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

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，$-\infty<a_i<\infty$；b_i 為試題 i 之試題難度參數，$-\infty<b_i<\infty$；c_i 為試題 i 之試題猜測度參數（item guessing parameter），$0\le c_i<1$。

三、一般化部分給分模式（GPCM）

Muraki（1992）所提出，為各試題之間有不同的鑑別度參數。GPCM模式假定一試題 j 具有 m_j 個等級類別（graded categories），越高的類別表示能力越高，而最高得分為 m_j，GPCM模式如下：

$$P_{jk}(\theta)=\frac{\exp\left[\sum_{v=1}^{k}a_j(\theta-b_{jv})\right]}{\sum_{c=1}^{m_j}\left[\exp\sum_{v=1}^{c}a_j(\theta-b_{jv})\right]}=\frac{\exp\left[\sum_{v=1}^{k}a_j(\theta-b_{jv}+d_v)\right]}{\sum_{c=1}^{m_j}\left[\exp\sum_{v=1}^{c}a_j(\theta-b_{jv}+d_v)\right]}$$

其中

θ：表示受試者的潛在能力特質（$-\infty<\theta<\infty$）。

k：為受試者的回答所屬類別，從1…m_j。

e：是底為1.728的指數。

m_j：為隨題目而變的變數，m_j 則是第 j 題所有的類別數。

$P_{jk}(\theta)$：為潛在能力特質為 θ 的受試者在第 j 題得到第 k 類的機率（$0<P_{jk}(\theta)<1$）。

b_{jv}：$b_{jv}=b_j-d_v$。b_{jv} 為第 j 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_{jv} 參數（$-\infty<b_{jv}<\infty$），即 b_{jk} 為 $P_{j,k-1}(\theta)$ 和 $P_{jk}(\theta)$ 的交點，同一試題內的試題步驟參數不需是有序的。b_j 為試題座標參數（item location parameter）、d_v 為閾參數（threshold parameter），d_k 為同一試題內的第 k 類和其他類別的相對難度（Andrich, 1982）。

a_j：試題 j 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。

四、多向度隨機係數多項洛基模式（MRCMLM）

MRCMLM是由Adams、Wilson與Wang（1997）等人所提出，MRCMLM為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ξ所描述，而受試者的潛在變數θ，是一個隨機變項，其模式定義如下：

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

其中，X_{ik}：受試者之做答反應組型

K_i：第i試題的計分類別數

θ：受試者的能力參數矩陣（多向度能力）

ξ：試題參數向量

$\mathbf{a}_{ik}$：第i題中第k個反應類別的設計向量（designing vector）

$\mathbf{b}_{ik}$：第i題在第k個反應類別上的計分向量（scoring vector）

A：整份測驗的設計矩陣（designing matrix）

B：整份測驗的計分矩陣（scoring matrix）

表4-2-2為整理各大型測驗中，不同的測驗題型所使用的模式。

表4-2-2　大型測驗所使用之測量模式

題型	NAEP	TIMSS	PISA	TASA
選擇題	三參數對數模式（Birnbaum, 1968; Lord, 1980）	三參數對數模式（Birnbaum, 1968; Lord, 1980）	多向度隨機係數多項洛基模式，MRCML（Adams, Wilson & Wang, 1997）	三參數對數模式（Birnbaum, 1968; Lord, 1980）
填充題	二參數對數模式（Birnbaum, 1968）	二參數對數模式（Birnbaum, 1968）		無此題型
開放性試題	一般化部分給分模式（Muraki, 1992）	一般化部分給分模式（Muraki, 1992）		敘述統計分析

五、綜合討論與建議

目前國際大型測驗在選擇題題型部分NAEP、TIMSS皆使用三參數對數模式，PISA則使用多向度隨機係數多項洛基模式，目前TASA為使用三參數對數模式，未來建議繼續使用此模式進行選擇題的分析，題組試題的部分，建議使用題組模式分析；開放性試題部分NAEP、TIMSS皆使用一般化部分給分模式，PISA則使用多向度隨機係數多項洛基模式，目前TASA針對開放性試題僅進行敘述統計分析，未來建議使用一般化部分給分模式進行試題分析。

參、模式適合度評估方法

在模式適合度方面，NAEP、TIMSS中的模式適合度是使用圖形化判斷方法，PISA是以標準殘差（standardised residual）為基礎，建立非權重的適合度統計量（unweighted fit statistic）和權重適合度統計量（weighted fit statistic）。詳述如下：

一、圖形化判斷方法

在NAEP、TIMSS中使用試題適合統計量，因為沒有一個真正的χ^2分佈，測量試題適合統計量為比較真實與理論上的試題反應函數，看試題對於模式而言是否較不合適，像是多點計分試題中某個類別的得分較低，或者某一題的反應與理論上不符合。對於IRT模式適合度的方法，為比較同一量尺上的觀察值以及理論上的試題反應函數所產生的曲線，其中理論上的曲線（theoretical curves）是根據試題參數的估計值所畫出來的，而觀察值則是依據有施測該試題的學生所產生的後驗分佈而得。對二元計分試題而言，能力值為θ的學生答對該試題的後驗分佈加上能力值為θ的學生遺漏該試題的後驗分佈，此方法相似於答對該題的學生加上遺漏該題的學生。在每個能力值上學生施測該試題（receiving the item）的後驗分佈機率值的總和，相似於在各能力點上施測該試題的學生數。最後的試題反應曲線的值（plotted values）為各能力值的個別後驗機率值的加總，利用答對該題的後驗機率值加上遺漏該題的後驗機率值除以施測該題的後驗機率值，在估計完試題參數後，通常是透過估算試題的適配程度代表IRT模式的適合度。利用圖形的分析做為模式適合度的評估準則，比較實際值與理論值的曲線來做為判斷式，兩者的圖形越接近重疊，則適配的情形越好。

二、適合度統計量

PISA中使用ConQuest 軟體所提供的適合度檢定方法，ConQuest 軟體針對每個需要估計的參數，提供一個適合度的檢定，此檢定是Wu（1997）以Wright 及 Masters（1982）所提出的論點為基礎所發展出來的，Wu（1997）將它延伸到兩個面向。第一，將它應用到更廣泛的模式中，提供參數的適合度檢定，而非原始的試題的適合度檢定。第二，Wright 和 Masters（1982）所提出之適合度統計方法適用於非條件式的最大概似估計法中（unconditional maximum likelihood estimates），而Wu（1997）將其延伸至可用至邊際最大概似估計法中（marginal maximum likelihood estimates）。

令A_p為設計矩陣A中的第p行，Wu（1997）的適合度統計是以標準殘差為基礎的。

$$z_{np}(\theta_n) = \frac{A'_p x_n - E_{np}}{\sqrt{V_{np}}}$$

其中，$A'_p x_n$ 為受試者 n 在參數 p 上的充份統計量，E_{np} 和 V_{np} 分別為 $A'_p x_n$ 的條件期望值與變異數，建立一個非權重的適合度檢定，殘差的平方為個別後驗機率分佈積分的平均。

$$Fit_{out,p} = \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_N}\left[\frac{1}{N}\sum_{n=1}^{N}\hat{z}_{np}^2(\theta_n)\right]\prod_{n=1}^{N}h_\theta(\theta_n;Y_n,\hat{\xi},\hat{\beta},\hat{\sigma}^2\mid x_n)d\theta_N d\theta_{N-1}\cdots d\theta_1$$

針對權重的適合度檢定，殘差平方的權重平均可以表示如下式所示：

$$Fit_{in,p} = \int_{\theta_1}\int_{\theta_2}\cdots\int_{\theta_N}\left[\frac{\sum_{n=1}^{N}\hat{z}_{np}^2(\theta_n)V_{np}(\theta_n)}{\sum_{n=1}^{N}V_{np}(\theta_n)}\right]\prod_{n=1}^{N}h_\theta(\theta_n;Y_n,\hat{\xi},\hat{\beta},\hat{\sigma}^2\mid x_n)d\theta_N d\theta_{N-1}\cdots d\theta_1$$

在ConQuest中，蒙地卡羅方法用來逼近上述方程式的積分，Wu（1997）表示上述方程式近似於卡方分配，利用Wilson-Hilferty transformations轉換方法將統計量轉換成近似於常態。

$$t_{out,p} = \frac{(Fit_{out,p}^{\frac{1}{3}} - 1 + \frac{2}{(9rN)})}{(\frac{2}{9rN})^{\frac{1}{2}}}$$

和

$$t_{in,p} = \left[Fit_{out,p}^{\frac{1}{3}} - 1\right]\times\frac{3}{\sqrt{Var(Fit_{in,p})}} + \frac{\sqrt{Var(Fit_{in,p})}}{3}$$

其中，r 為蒙地卡羅法的抽取次數。

$$Var(Fit_{in,p}) = \left[\frac{1}{\sum_n V_{np}}\right]^2\times\frac{3}{\sqrt{Var(Fit_{in,p})}} + \frac{\sqrt{Var(Fit_{in,p})}}{3}$$

詳細的推導過程請詳閱Wu（1997）。表4-2-3是各大型測驗中模式適合度評估之方法。

表4-2-3　模式適合度評估方法

NAEP	TIMSS	PISA	TASA
圖形化判斷	圖形化判斷	Unweighted fit statistic and Weighted fit statistic	無

三、綜合討論與建議

模式適合度評估方法目前NAEP、TIMSS皆使用圖形化判斷方式進行評估，PISA則使用ConQuest 軟體所提供的適合度檢定方法，Unweighted fit statistic and Weighted fit statistic，但由於此方法僅適用於Rasch模式下，因此不適合目前為使用三參數對數模式的TASA，TASA目前尚未使用任何方法進行模式適合度評估，未來建議參考NAEP、TIMSS使用圖形化判斷方式進行模式適合度評估。

肆、測驗分析軟體

不同的大型測驗使用不同的分析軟體進行參數之估計，TIMSS中分別使用BILOG進行二參數對數模式、三參數對數模式試題進行分析，使用PARSCALE進行一般化部分給分模式試題進行分析；NAEP中使用結合BILOG和PARSCALE的NAEP BILOG/PARSCALE軟體進行二參數對數模式、三參數對數模式以及一般化部分給分模式試題分析；PISA中使用的模式為多向度隨機係數多項洛基模式，因此其參數估計軟體為使用適合多向度隨機係數多項洛基模式的ConQuest進行分析，各大型測驗所使用的估計軟體整理如表4-2-4所示。

表4-2-4　大型測驗使用之測驗分析軟體

NAEP	TIMSS	PISA	TASA
BILOG（Mislevy & Bock's, 1982） PARSCALE（Muraki & Bock's, 1991）	BILOG（Mislevy & Bock's, 1982） PARSCALE（Muraki & Bock's, 1991）	ConQuest（Wu, Adams, & Wilson, 1998）	BILOG（Mislevy & Bock's, 1982） SCORIGHT（Wang, Bradlow & Wainer, 2004）

一、綜合討論與建議

目前大型測驗中NAEP、TIMSS依題型選擇題、填充題以及開放性試題使用分析模式分別為三參數對數模式、二參數對數模式和一般化部分給分模式，因此分別使用BILOG、PARSCALE軟體進行分析，PISA則是使用多向度隨機係數多項洛基模式進行三種題型的試題分析，因此選用ConQuest，目前TASA針對選擇題部分使用BILOG進行分析，未來建議TASA繼續使用BILOG進行選擇題分析，並增加PARSCALE進行開放性試題分析。

第三節　試題特性

試題特性分析對於國內外大型測驗評比資料庫是資料處理上相當重要的一環，藉由測量模式或一般性描述統計分析試題各項參數的穩定性可以提高測驗的品質，確實瞭解學生學習成就表現的變化。本節探討內容重點在綜合討論比較NAEP、PISA、TIMSS三個國際大型測驗評比資料庫與TASA在試題特性分析上之差異，並提出具體建議，期望整理出一套普遍一致的標準化測驗試題分析流程，供TASA後續研究分析使用。

NAEP、PISA、TIMSS等國際大型測驗評比資料庫為了確保測驗的信度及效度，對於測驗過程中所使用的試題皆會進行不同的統計分析及信、效度檢測，其中包含了一般性試題統計描述、試題參數估計、評分者一致性效應檢測及差異試題功能分析（DIF）等內容。同樣的，TASA身為國內第一個大型測驗評比資料庫，對於測驗中所使用的試題依然會進行分析、檢測與篩選。底下即分別針對NAEP、PISA、TIMSS及TASA資料庫對於試題特性所採用的分析方法進行比較，期望獲得普遍一致的標準化分析流程，以提升TASA測驗的信度及效度。

壹、試題統計描述

一、NAEP

分為二元計分試題及多元計分試題兩種，所呈現的相關數據如下：

（一）二元計分試題（dichotomously scored items）

人數、選項百分比、答對率、遺漏未作答百分比、無效試卷百分比、答對率、難度指標轉換為平均=13、標準差=4、二系列相關、點二系列相關。

（二）多元計分試題（polytomously scored items）

無效試卷百分比、難度指標轉換為平均=13、標準差=4、多系列相關、Pearson相關。

二、PISA

（一）國家報告呈現格式

1. 以表格的形式呈現個別試題統計描述

表4-3-1為PISA 2006技術報告有關試題描述性統計分析之表格呈現方式，第一行標籤部分，1, 2, 3, 4為作答反應選項，8為無效值，9為遺漏值，其餘分析包含了人數、鑑別度、閾值、試題不合適之均方根的值（item infit mean square）、選項百分比、答對率、遺漏未作答百分比、點二系列相關（t檢定）、PV1 Avg:1 與PV1 SD:1等等。

表4-3-1　試題之描述性統計分析（表格化）

Example of item statistics in Report 1

Item:70 (S423Q01)
Cases for this item　355　Discrimination　0.13
Item Threshold(s):　0.49　Weighted MNSQ　1.17
Item Delta(s):　0.49

Label	Score	Count	% of tot	Pt Bis	t (p)	PV1Avg:1	PV1 SD:1
0		0	0.00	NA	NA (.000)	NA	NA
1	0.00	65	18.31	-0.16	-3.02(.003)	-0.30	0.78
2	1.00	138	38.87	0.13	2.54(.011)	0.12	0.89
3	0.00	115	32.39	0.09	1.76(.080)	0.07	0.83
4	0.00	26	7.32	-0.08	-1.44(.152)	-0.41	0.83
5		0	0.00	NA	NA (.000)	NA	NA
6		0	0.00	NA	NA (.000)	NA	NA
8	0.00	4	1.13	-0.06	-1.19(.233)	-0.62	0.79
9	0.00	7	1.97	-0.15	-2.87(.004)	-0.76	0.58

註：PV1 Avg:1 和PV1 SD:1 是學生反應在每一類別的平均，以上表而言，正確答對的學生的平均可能值是0.12，作答反應1之學生的平均可能值是-0.3，依此類推。

資料來源：PISA 2006 Technical Report, p.148

2. 以圖的形式呈現個別試題統計描述

Example of item statistics in Report 2

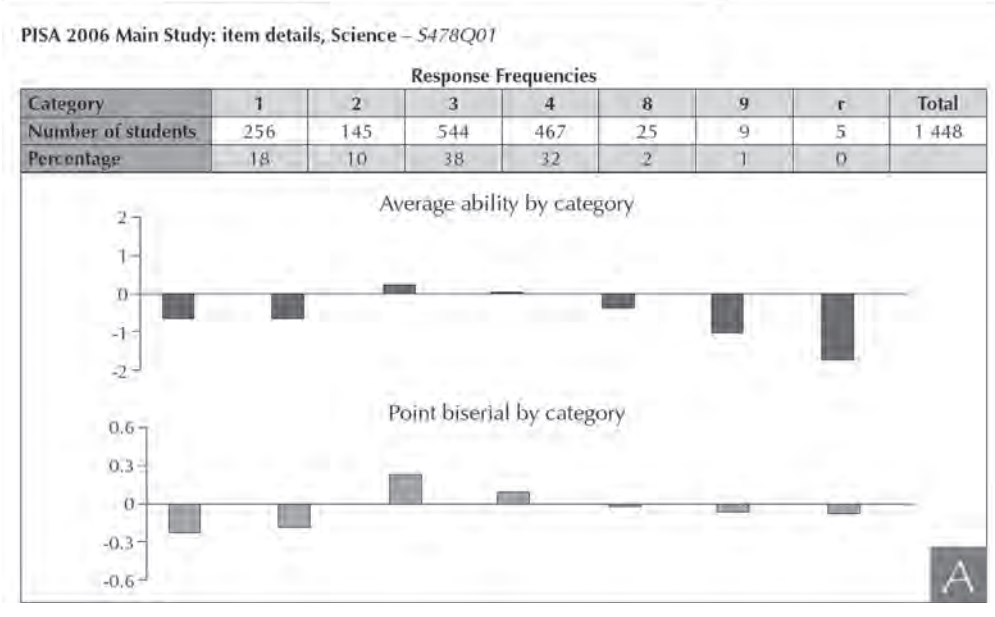


圖4-3-1　試題選項之平均能力值與點二系列相關值

資料來源：PISA 2006 Technical Report, p.149

3. 試題在國家與國際間之模式適合度與鑑別度參數比較

Example of item statistics shown in Graph B

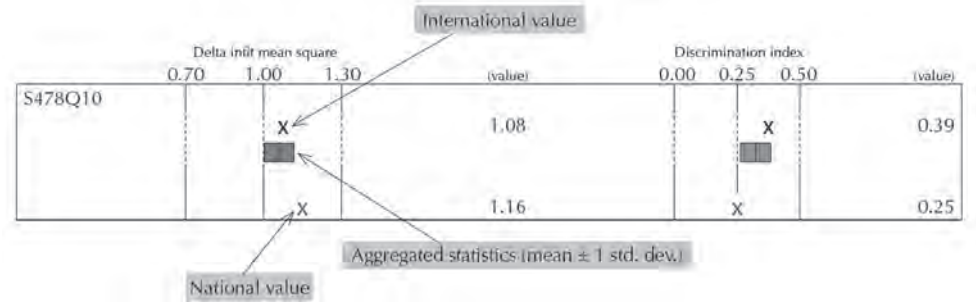


圖4-3-2　試題在國家與國際間之模式適合度與鑑別度參數比較
資料來源：PISA 2006 Technical Report, p.150

4. 試題在國家與國際間之難度與閾值比較

Example of item statistics shown in Graph C

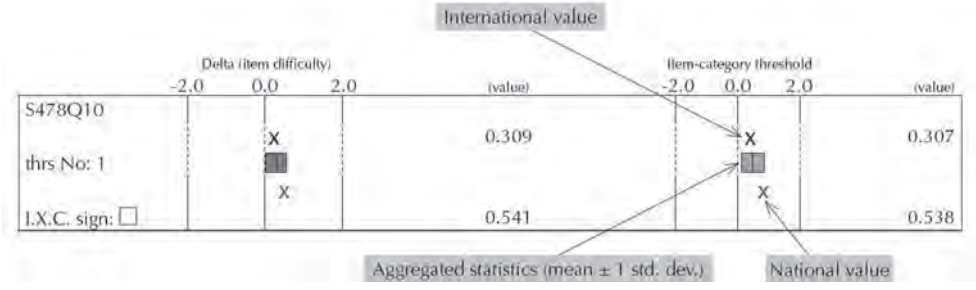


圖4-3-3　試題在國家與國際間之難度與閾值比較
資料來源：PISA 2006 Technical Report, p.151

5. 期望分數曲線

觀察試題在單一國家的實際表現、國際間表現與預期分數表現之差異，如圖4-3-4。

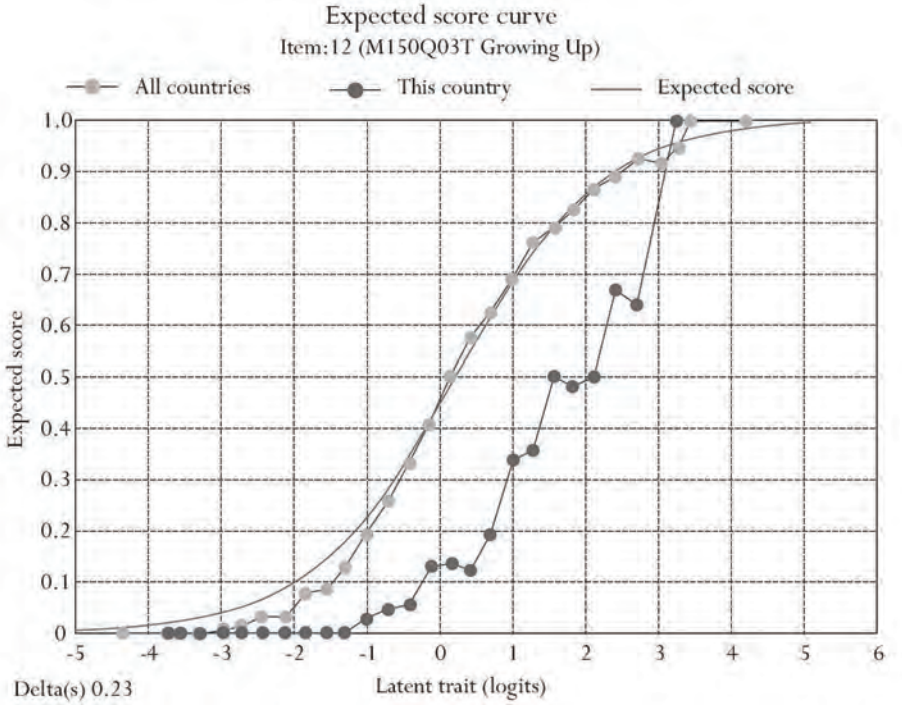


圖4-3-4　單一試題在單一國家、國際間與期望之分數表現曲線比較

資料來源：PISA 2003 Technical Report, p.127

三、TIMSS

分為選擇題及開放性試題兩種，另外針對定錨試題會呈現年度間的數據比較，並利用試題難度值的在各國間的不同檢視是否存在試題效應，所呈現的相關數據如下：

（一）選擇題

表4-3-2為人數、答對率（古典難度）、鑑別度（點二系列相關）、選項百分比、缺失遺漏未作答百分比、相關、單參難度。

Trends in International Mathematics and Science Study - TIMSS 2007 Assessment Results
International Item Statistics (Unweighted) - Review Version - 8th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics: Algebra / Knowing (M042077 - M10_06)
Label: Expression to equivalent to 4(3+x)
Type: MC Key: C

				Percentages							Point Biserials							
Country	N	Diff	Disc	Pct_A	Pct_B	Pct_C	Pct_D	Pct_E	Pct_OM	Pct_NR	PB_A	PB_B	PB_C	PB_D	PB_E	PB_OM	RDIFF	Flags
Algeria	752	16.2	0.11	40.2	19.8	16.2	19.3	.	4.5	2.0	0.03	-0.14	0.11	0.01	.	-0.06	0.36	DCH
Armenia	683	73.2	0.46	15.4	5.0	73.2	3.7	.	2.8	0.4	-0.29	-0.19	0.45	-0.19	.	-0.08	-1.01	F
Australia	578	31.1	0.40	32.0	19.0	31.1	15.6	.	2.2	0.3	-0.11	-0.28	0.40	-0.01	.	-0.14	0.97	H
Bahrain	597	48.4	0.47	22.1	13.4	48.4	14.9	.	1.2	0.2	-0.15	-0.31	0.47	-0.16	.	-0.10	-1.11	E
Bosnia and Herzegovin	606	58.1	0.51	22.9	5.8	58.1	11.1	.	2.1	0.0	-0.30	-0.28	0.51	-0.14	.	-0.13	-0.97	F
Botswana	586	24.2	0.36	30.9	28.3	24.2	15.4	.	1.2	2.0	0.04	-0.30	0.36	-0.10	.	-0.05	-0.13	CH
Bulgaria	576	73.6	0.55	14.8	5.7	73.6	3.8	.	2.1	0.9	-0.32	-0.29	0.55	-0.21	.	-0.12	-1.44	E F
Chinese Taipei	572	85.8	0.61	7.5	2.3	85.8	3.7	.	0.7	0.0	-0.42	-0.30	0.61	-0.28	.	-0.06	-1.05	F
Colombia	682	26.0	0.25	36.2	15.1	26.0	21.1	.	1.6	2.7	-0.14	-0.17	0.25	0.06	.	-0.06	-0.29	
Cyprus	622	56.9	0.47	22.0	7.6	56.9	12.2	.	1.3	0.0	-0.27	-0.28	0.47	-0.11	.	-0.11	-0.73	F
Czech Republic	693	59.2	0.49	20.3	5.2	59.2	15.2	.	0.1	0.0	-0.33	-0.20	0.49	-0.18	.	-0.05	-0.38	F
Egypt	945	56.9	0.42	19.3	11.9	56.9	11.3	.	0.6	0.0	-0.16	-0.29	0.42	-0.15	.	-0.07	-1.28	E
El Salvador	568	24.6	0.32	34.9	19.9	24.6	19.0	.	1.6	1.4	-0.10	-0.22	0.32	0.01	.	-0.07	-0.60	C
England	571	46.1	0.59	20.8	18.4	46.1	13.7	.	1.1	0.0	-0.16	-0.46	0.59	-0.13	.	-0.09	0.53	H
Georgia	616	55.7	0.51	23.9	7.1	55.7	11.2	.	2.1	1.1	-0.30	-0.22	0.51	-0.16	.	-0.09	-1.19	E F
Ghana	751	42.1	0.46	24.5	18.0	42.1	13.8	.	1.6	1.1	-0.14	-0.27	0.46	-0.12	.	-0.08	-1.45	E
Hong Kong SAR	502	87.5	0.51	5.6	2.8	87.5	3.8	.	0.4	0.0	-0.37	-0.32	0.51	-0.13	.	-0.10	-1.39	E F
Hungary	599	63.4	0.60	15.7	6.5	63.4	13.4	.	1.0	0.2	-0.30	-0.30	0.60	-0.26	.	-0.13	-0.45	F
Indonesia	591	36.9	0.47	28.4	13.7	36.9	20.3	.	0.7	0.5	-0.16	-0.26	0.47	-0.14	.	0.02	-0.50	
Iran, Islamic Rep. of	574	40.9	0.45	30.7	7.1	40.9	20.7	.	0.5	0.2	-0.18	-0.18	0.45	-0.21	.	-0.09	-0.61	F
Israel	453	71.3	0.52	14.8	4.0	71.3	9.3	.	0.7	1.5	-0.27	-0.26	0.52	-0.22	.	-0.14	-1.48	E F
Italy	626	48.1	0.38	26.0	6.9	48.1	17.9	.	1.1	0.5	-0.19	-0.23	0.38	-0.10	.	-0.07	-0.21	H F
Jordan	761	53.9	0.58	18.3	14.5	53.9	12.5	.	0.9	0.1	-0.19	-0.37	0.58	-0.23	.	-0.10	-1.02	E
Korea, Rep. of	603	85.6	0.58	7.5	3.6	85.6	3.3	.	0.0	0.0	-0.37	-0.32	0.58	-0.25	.	0.00	-0.89	F
Kuwait	571	23.5	0.32	31.2	20.8	23.5	21.9	.	2.6	0.7	0.00	-0.23	0.32	-0.07	.	-0.10	-0.36	C
Lebanon	538	79.0	0.38	11.7	2.0	79.0	6.5	.	0.7	0.6	-0.24	-0.13	0.38	-0.19	.	-0.13	-2.04	E F
Lithuania	571	62.7	0.59	16.1	7.7	62.7	12.6	.	0.9	0.0	-0.31	-0.32	0.59	-0.24	.	-0.07	-0.52	F
Malaysia	634	44.6	0.50	27.3	13.1	44.6	14.4	.	0.6	0.2	-0.15	-0.35	0.50	-0.15	.	-0.12	-0.10	H
Malta	671	60.4	0.51	18.8	8.9	60.4	10.9	.	1.0	0.1	-0.17	-0.35	0.51	-0.22	.	-0.14	-0.65	F
Mongolia	610	57.4	0.49	18.7	7.9	57.4	14.4	.	1.6	0.3	-0.28	-0.25	0.49	-0.20	.	0.00	-1.38	E F
Morocco	614	40.9	0.42	27.9	15.5	40.9	14.3	.	1.5	0.5	-0.17	-0.31	0.42	-0.05	.	-0.05	-0.94	
Norway	656	15.5	0.15	36.1	29.3	15.5	15.2	.	3.8	0.3	0.10	-0.18	0.15	0.01	.	-0.14	1.54	CH
Oman	678	41.7	0.49	20.5	20.8	41.7	16.1	.	0.9	0.3	-0.09	-0.34	0.49	-0.17	.	-0.03	-1.08	E
Palestinian Nat'l Aut	653	37.1	0.44	23.6	22.4	37.1	15.3	.	1.7	0.3	-0.03	-0.34	0.44	-0.14	.	-0.08	-0.71	
Qatar	1017	31.7	0.26	27.7	23.1	31.7	15.7	.	1.8	0.1	-0.02	-0.20	0.26	-0.06	.	-0.04	-1.07	E
Romania	601	65.6	0.53	20.6	5.8	65.6	7.5	.	0.5	0.3	-0.33	-0.25	0.52	-0.17	.	-0.10	-1.17	E F
Russian Federation	633	77.3	0.52	12.6	2.8	77.3	6.2	.	1.1	0.8	-0.32	-0.15	0.52	-0.29	.	-0.12	-1.27	E F
Saudi Arabia	612	22.2	0.31	30.2	23.0	22.2	23.4	.	1.1	0.2	0.03	-0.24	0.31	-0.07	.	-0.06	-0.47	C
Scotland	564	43.1	0.54	19.3	22.2	43.1	13.8	.	1.6	0.2	-0.13	-0.38	0.54	-0.13	.	-0.10	0.20	H
Serbia	572	69.9	0.61	16.3	4.0	69.9	8.6	.	1.2	0.2	-0.37	-0.26	0.61	-0.26	.	-0.15	-1.22	E F
Singapore	655	86.3	0.58	8.1	1.8	86.3	3.5	.	0.3	0.0	-0.38	-0.28	0.58	-0.28	.	-0.12	-1.13	F
Slovenia	570	37.5	0.41	36.5	4.4	37.5	21.1	.	0.5	0.0	-0.25	-0.18	0.41	-0.09	.	-0.06	0.64	H F
Sweden	725	24.6	0.20	36.6	24.1	24.6	11.6	.	3.2	0.5	0.08	-0.23	0.20	0.03	.	-0.18	1.36	CH
Syrian Arab Republic	670	46.1	0.45	23.3	13.9	46.1	15.4	.	1.3	0.3	-0.18	-0.24	0.45	-0.13	.	-0.11	-1.02	E
Thailand	763	37.0	0.55	27.0	18.1	37.0	16.9	.	1.0	0.4	-0.14	-0.34	0.55	-0.16	.	-0.10	0.00	H
Tunisia	580	34.8	0.48	14.5	31.2	34.8	18.4	.	1.0	0.3	-0.09	-0.35	0.48	-0.07	.	-0.07	-0.25	
Turkey	663	53.5	0.60	21.0	14.9	53.5	9.8	.	0.8	0.2	-0.30	-0.34	0.60	-0.18	.	-0.04	-0.91	F
Ukraine	640	66.4	0.51	16.7	6.7	66.4	9.4	.	0.8	0.2	-0.26	-0.29	0.51	-0.22	.	-0.08	-1.35	E F
United States	1049	50.4	0.53	28.5	4.9	50.4	15.3	.	0.9	0.2	-0.29	-0.24	0.53	-0.19	.	-0.08	0.14	H F
International Avg.	.	50.5	0.46	22.6	12.5	50.5	13.1	.	1.3	0.5	-0.19	-0.27	0.46	-0.14	.	-0.09	-0.61	
Basque Country, Spain	331	66.5	0.40	18.7	2.1	66.5	11.2	.	1.5	0.0	-0.26	-0.19	0.40	-0.12	.	-0.16	-0.83	F
British Columbia, Can	600	34.8	0.42	41.7	8.0	34.8	13.7	.	1.8	0.3	-0.21	-0.19	0.42	-0.10	.	-0.10	0.88	H F
Massachusetts, US	282	56.0	0.58	25.2	2.8	56.0	14.5	.	1.4	0.0	-0.35	-0.20	0.58	-0.24	.	-0.14	0.39	H F
Minnesota, US	247	42.1	0.52	35.6	5.7	42.1	16.2	.	0.4	0.0	-0.26	-0.23	0.52	-0.20	.	-0.11	0.79	H F
Ontario, Canada	496	21.4	0.26	44.6	11.3	21.4	20.2	.	2.6	0.4	-0.02	-0.24	0.26	-0.01	.	-0.09	1.88	CH
Quebec, Canada	590	56.1	0.44	24.6	5.4	56.1	12.7	.	1.2	0.7	-0.20	-0.24	0.44	-0.19	.	-0.11	0.01	H F

Keys: Diff: Percent correct score; Disc: Item discrimination; Pct_A...E: Percent choosing option; Pct_OM, NR: Percent Omitted and Not Reached; PB_A...E: Point Biserial for option; PB_OM: Point Biserial for Omittted. RDIFF= Rasch difficulty.

Flags: A= Ability not ordered/Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average; F= Distractor chosen by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

（二）開放性試題

表4-3-3呈現人數、答對率（古典難度）、鑑別度（點二系列相關）、選項百分比、遺漏未作答百分比、相關、1PL之難度。

Trends in International Mathematics and Science Study - TIMSS 2007 Assessment Results
International Item Statistics (Unweighted) - Review Version - 8th Grade
For Internal Review Only: DO NOT CITE OR CIRCULATE

Mathematics: Number / Applying (M042304B - M04_05B)
Label: Complete the table for cycling
Type: CR Key: X

				Percentages					Point Biserials					Reliability			
Country	N	Diff	Disc	Pct_0	Pct_1	Pct_2	Pct_OM	Pct_NR	PB_0	PB_1	PB_2	PB_OM	RDIFF	Cases	Score	Code	Flags
Algeria	680	12.8	0.26	57.6	22.1	1.8	18.5	1.6	-0.18	0.10	0.32	-0.02	0.49	188	60.1	57.4	FR
Armenia	676	15.3	0.54	39.8	3.4	13.6	43.2	0.4	-0.01	0.18	0.50	-0.39	1.63	219	98.2	96.3	H F
Australia	571	38.2	0.60	33.5	37.1	19.6	9.8	0.3	-0.35	0.14	0.49	-0.32	0.35	190	96.8	96.8	E
Bahrain	600	15.4	0.50	58.2	24.2	3.3	14.3	0.5	-0.32	0.30	0.37	-0.11	0.80	200	98.5	98.5	F
Bosnia and Herzegovin	600	19.8	0.48	32.2	27.7	6.0	34.2	0.0	-0.12	0.24	0.38	-0.30	1.09	210	90.5	89.5	H F
Botswana	596	6.2	0.54	74.0	9.4	1.5	15.1	0.3	-0.24	0.35	0.41	-0.13	1.12	207	96.6	96.6	F
Bulgaria	569	22.8	0.60	36.4	21.6	12.0	30.1	0.4	-0.27	0.25	0.49	-0.29	1.25	182	80.2	79.1	H
Chinese Taipei	581	57.7	0.65	20.8	31.5	42.0	5.7	0.0	-0.44	0.00	0.53	-0.37	1.00	219	98.2	98.2	
Colombia	701	24.6	0.58	46.1	32.4	8.4	13.1	2.0	-0.39	0.25	0.46	-0.13	-0.22	197	98.5	98.0	E F
Cyprus	635	34.8	0.53	21.1	39.7	15.0	24.3	0.0	-0.17	0.11	0.45	-0.34	0.22	.	.	.	E
Czech Republic	693	43.3	0.57	23.2	42.1	22.2	12.4	0.1	-0.31	0.03	0.49	-0.27	0.10	200	96.5	96.0	E
Egypt	955	15.7	0.64	65.3	18.8	6.3	9.5	0.2	-0.46	0.38	0.46	-0.14	0.88	203	100.0	100.0	F
El Salvador	569	10.2	0.43	63.4	17.2	1.6	17.8	2.2	-0.27	0.30	0.31	-0.08	0.56	203	100.0	100.0	F
England	560	43.0	0.63	25.2	41.8	22.1	10.9	0.0	-0.37	0.10	0.51	-0.32	0.49	84	98.8	98.8	
Georgia	642	15.1	0.49	32.7	19.3	5.5	42.5	0.8	-0.12	0.33	0.32	-0.31	0.87	212	91.0	89.2	F
Ghana	754	3.7	0.38	70.0	6.4	0.5	23.1	0.4	-0.11	0.32	0.19	-0.10	1.35	269	99.3	99.3	H F
Hong Kong SAR	484	50.2	0.60	25.4	38.4	31.0	5.2	0.4	-0.39	0.05	0.48	-0.34	1.07	164	95.1	93.9	H
Hungary	584	49.4	0.64	17.5	44.3	27.2	11.0	0.0	-0.36	-0.05	0.56	-0.29	0.21	181	94.5	92.8	E
Indonesia	596	11.1	0.55	63.8	13.4	4.4	18.5	0.2	-0.27	0.34	0.40	-0.17	0.88	203	93.1	92.6	F
Iran, Islamic Rep. of	568	17.3	0.53	54.0	24.6	4.9	16.4	0.4	-0.32	0.29	0.40	-0.16	0.71	182	98.4	97.8	F
Israel	485	28.5	0.57	36.5	32.2	12.4	19.0	0.8	-0.31	0.30	0.40	-0.31	0.63	239	90.8	85.4	
Italy	620	32.8	0.47	23.9	41.5	12.1	22.6	0.2	-0.22	0.14	0.39	-0.25	0.58	216	99.5	99.5	
Jordan	752	14.4	0.52	68.8	18.8	5.1	7.4	0.1	-0.37	0.33	0.37	-0.14	1.10	221	98.6	98.6	H F
Korea, Rep. of	601	53.7	0.71	24.5	33.6	36.9	5.0	0.0	-0.47	-0.09	0.64	-0.29	1.22	212	97.2	97.2	H
Kuwait	587	8.7	0.54	70.2	11.2	3.1	15.5	0.5	-0.22	0.31	0.43	-0.20	0.33	200	99.0	99.0	F
Lebanon	549	23.8	0.59	39.0	34.8	6.4	19.9	0.2	-0.37	0.38	0.36	-0.21	1.03	191	99.5	98.4	F
Lithuania	577	37.8	0.59	30.0	34.3	20.6	15.1	0.0	-0.32	0.07	0.52	-0.27	0.58	174	96.6	95.4	
Malaysia	627	32.5	0.61	43.1	33.3	15.8	7.8	0.0	-0.46	0.21	0.48	-0.16	0.43	216	96.3	96.3	
Malta	660	42.7	0.63	29.7	34.2	25.6	10.5	0.2	-0.39	0.10	0.52	-0.32	0.18	176	93.8	93.8	E
Mongolia	631	15.3	0.35	42.2	24.2	3.2	30.4	0.6	0.05	0.17	0.32	-0.34	1.18	192	81.3	72.9	H F
Morocco	615	20.7	0.22	42.3	33.7	3.9	20.2	0.5	0.00	-0.05	0.37	-0.13	0.39	176	93.2	93.2	AF
Norway	656	34.0	0.54	28.7	39.9	14.0	17.4	0.5	-0.24	0.18	0.42	-0.33	0.07	220	94.1	92.3	E
Oman	675	7.0	0.53	80.0	13.0	0.4	6.5	0.9	-0.39	0.51	0.14	-0.09	1.93	226	99.1	99.1	H F
Palestinian Nat'l Aut	642	9.7	0.55	75.4	13.2	3.1	8.3	0.0	-0.41	0.42	0.33	-0.09	0.83	231	94.4	94.4	F
Qatar	1019	5.0	0.50	71.9	8.4	0.8	18.8	0.1	-0.16	0.40	0.29	-0.16	0.98	197	98.0	97.5	F
Romania	599	31.1	0.60	30.4	32.2	15.0	22.4	0.3	-0.29	0.14	0.53	-0.28	0.44	209	98.6	97.6	
Russian Federation	637	42.5	0.60	22.6	39.1	22.9	15.4	0.3	-0.32	0.02	0.54	-0.28	0.58	212	99.1	99.1	
Saudi Arabia	600	5.0	0.36	79.0	8.3	0.8	11.8	0.7	-0.14	0.29	0.22	-0.14	0.91	211	99.1	99.1	F
Scotland	585	37.5	0.62	27.2	41.9	16.6	14.4	0.2	-0.32	0.14	0.50	-0.34	0.37	116	99.1	98.3	E
Serbia	584	33.1	0.54	19.7	34.4	15.9	30.0	0.2	-0.15	0.14	0.45	-0.38	0.65	206	97.6	97.1	
Singapore	665	63.7	0.71	19.1	27.2	50.1	3.6	0.0	-0.50	-0.14	0.64	-0.32	0.52	220	99.1	99.1	
Slovenia	559	41.9	0.54	24.0	44.7	19.5	11.8	0.0	-0.26	-0.04	0.52	-0.23	0.25	184	100.0	100.0	E
Sweden	751	41.2	0.64	22.6	34.0	24.2	19.2	0.3	-0.25	0.04	0.56	-0.40	-0.03	253	96.0	94.9	E
Syrian Arab Republic	668	17.0	0.38	58.1	25.6	4.2	12.1	0.0	-0.24	0.14	0.36	-0.06	0.50	195	99.5	98.5	F
Thailand	768	32.9	0.59	39.1	32.6	16.7	11.7	0.1	-0.37	0.16	0.49	-0.23	0.11	197	95.4	94.9	E
Tunisia	577	19.2	0.49	57.4	23.4	7.5	11.8	0.3	-0.38	0.24	0.38	-0.05	0.30	205	87.3	87.3	E F
Turkey	654	20.0	0.55	50.2	29.8	5.0	15.0	0.0	-0.32	0.26	0.47	-0.17	1.18	184	97.8	97.8	H F
Ukraine	646	18.7	0.43	46.1	20.4	8.5	24.9	0.0	-0.12	0.18	0.36	-0.26	1.13	212	85.8	84.9	H F
United States	1066	41.3	0.62	31.8	40.8	20.9	6.5	0.6	-0.40	0.09	0.52	-0.26	0.50	284	93.3	92.3	
International Avg.	.	26.9	0.54	42.7	27.7	13.1	16.5	0.4	-0.28	0.18	0.43	-0.23	0.69	.	95.1	94.3	
Basque Country, Spain	326	38.5	0.50	23.9	45.1	16.0	15.0	0.0	-0.20	0.00	0.48	-0.26	0.28	199	99.0	99.0	E
British Columbia, Can	603	40.2	0.53	28.9	42.3	19.1	9.8	0.2	-0.25	0.07	0.45	-0.32	0.49	85	95.3	94.1	
Massachusetts, US	262	51.9	0.64	23.3	41.2	31.3	4.2	0.8	-0.46	0.00	0.53	-0.24	0.54	68	95.6	95.6	
Minnesota, US	257	42.4	0.59	31.5	45.9	19.5	3.1	0.0	-0.47	0.08	0.48	-0.08	0.69	69	94.2	92.8	
Ontario, Canada	492	43.2	0.57	28.5	40.4	23.0	8.1	1.4	-0.33	0.07	0.47	-0.33	0.40	127	96.9	96.1	
Quebec, Canada	569	51.0	0.61	19.2	43.9	29.0	7.9	0.9	-0.31	-0.07	0.55	-0.31	0.21	183	95.1	95.1	E

Keys: Diff: Percent correct score; Disc: Item discrimination; Pct_0...3: Percent obtaining score level; Pct_OM, NR: Percent Omitted and Not Reached;
PB_0...3: Point Biserial for score level; PB_OM: Point Biserial for Omitted; RDIFF= Rasch difficulty;
Reliability: (Cases) Responses double scored; (Score) Percent agreement on score; (Code) Percent agreement on code.
Flags: A= Ability not ordered/Attractive distractor; C= Difficulty less than chance; D= Negative/low discrimination; E= Easier than average;
F= Score obtained by less than 10%; H= Harder than average; R= Scoring reliability < 80%; V= Difficulty greater than 95.

（三）定錨試題

表4-3-4為針對歷年等化過程中所使用的定錨試題，呈現年度、人數、選項答對百分比、答對率（古典難度）、無效未作答遺漏百分比、女生答對率、男生答對率等等相關數據。

表4-3-4　TIMSS 2003及TIMSS 2007試題參數比較（定錨題）

COUNTRY	Year	N	10	70	71	79	99	V1	NOT REACHED	OMIT	1.GIRL % Right	2.BOY % Right
Armenia	2003	464	33.2	18.6	16.8	11.7	19.6	33.2	0.8	18.9	33.0	33.5
	2007	273	48.7	8.8	21.2	1.8	19.4	48.7	0.7	18.7	48.1	49.1
Australia	2003	375	64.0	17.5	11.0	5.5	2.1	64.0	0.0	2.1	65.3	62.7
	2007	291	64.6	17.2	14.1	3.4	0.7	64.6	0.0	0.7	63.5	65.6
Chinese Taipei	2003	380	76.5	3.3	18.7	0.3	1.3	76.5	0.0	1.3	76.5	76.4
	2007	300	64.0	6.7	26.0	3.0	0.3	64.0	0.0	0.3	63.3	64.6
England	2003	291	64.2	13.0	15.4	6.1	1.2	64.2	0.0	1.2	63.8	64.8
	2007	305	69.2	14.8	10.8	4.6	0.7	69.2	0.0	0.7	70.8	67.5
Hong Kong SAR	2003	373	69.8	7.9	18.9	3.0	0.4	69.8	0.0	0.4	72.5	67.7
	2007	268	69.8	5.2	22.0	1.9	1.1	69.8	0.0	1.1	69.9	69.6
Hungary	2003	268	70.5	15.4	9.9	3.6	0.6	70.5	0.0	0.6	75.9	65.1
	2007	288	72.2	11.1	11.8	3.8	1.0	72.2	0.0	1.0	73.5	71.1
Iran, Islamic Rep. of	2003	352	41.7	22.8	20.3	6.7	8.5	41.7	0.0	8.5	48.2	37.7
	2007	274	39.4	25.9	20.1	8.4	6.2	39.4	0.0	6.2	36.9	41.4
Italy	2003	353	72.4	9.2	14.3	3.7	0.5	72.4	0.0	0.5	75.4	69.8
	2007	323	66.3	10.5	15.2	3.1	5.0	66.3	0.0	5.0	65.2	67.0
Latvia	2003	295	61.0	19.5	13.5	4.8	1.1	61.0	0.0	1.1	62.0	60.1
	2007	277	68.2	15.2	11.6	3.6	1.4	68.2	0.0	1.4	71.1	65.5
Lithuania	2003	371	58.3	23.2	9.2	7.3	1.9	58.3	0.0	1.9	59.3	56.3
	2007	285	54.7	22.8	14.4	7.0	1.1	54.7	0.0	1.1	54.7	54.8
Morocco	2003	339	14.8	32.1	12.5	28.2	12.5	14.8	0.0	12.5	10.2	18.9
	2007	300	12.0	16.7	7.7	47.7	16.0	12.0	0.7	15.3	10.5	13.7
Netherlands	2003	242	54.7	15.4	19.9	10.0	0.0	54.7	0.0	0.0	57.3	51.8
	2007	237	63.3	15.2	14.8	5.5	1.3	63.3	0.0	1.3	62.5	64.1
New Zealand	2003	354	61.1	17.8	14.2	4.9	2.1	61.1	0.0	2.1	59.9	62.3
	2007	349	56.7	18.3	15.8	8.3	0.9	56.7	0.0	0.9	56.3	57.2
Norway	2003	361	58.6	20.5	12.8	6.3	1.8	58.6	0.0	1.8	61.7	55.9
	2007	290	55.9	16.9	20.7	4.1	2.4	55.9	0.0	2.4	57.4	54.5
Russian Federation	2003	325	53.9	16.6	8.3	17.1	4.2	53.9	0.0	4.2	51.7	56.1
	2007	323	62.8	15.5	11.5	7.4	2.8	62.8	0.0	2.8	60.1	65.0
Scotland	2003	330	54.5	19.1	17.3	6.8	2.3	54.5	0.4	1.9	54.2	54.9
	2007	286	59.1	21.7	10.1	7.7	1.4	59.1	0.0	1.4	50.0	66.3
Singapore	2003	562	78.1	9.7	9.1	2.4	0.7	78.1	0.0	0.7	80.7	75.5
	2007	360	78.3	10.3	9.4	1.7	0.3	78.3	0.0	0.3	81.4	75.1
Slovenia	2003	280	58.9	18.4	13.6	6.4	2.7	58.9	0.0	2.7	63.2	
	2007	316	66.5	14.6	9.5	8.5	0.9	66.5	0.0	0.9	69.7	
Tunisia	2003	354	18.5	26.2	20.3	27.2	7.8	18.5	0.0	7.8	16.4	
	2007	290	18.6	29.3	13.1	30.0	9.0	18.6	0.0	9.0	20.7	
United States	2003	809	60.7	20.0	11.5	6.7	1.0	60.7	0.0	1.0	62.9	
	2007	566	61.3	19.3	12.0	6.0	1.4	61.3	0.2	1.2	59.2	
International Avg.	2003	.	56.3	17.3	14.4	8.4	3.6	56.3	0.1	3.6	57.5	
	2007	.	57.6	15.8	14.6	8.4	3.7	57.6	0.1	3.6	57.2	
Ontario, Canada	2003	359	55.1	21.3	14.6	7.1	2.0	55.1	0.0	2.0	49.1	
	2007	255	50.6	22.7	12.9	8.6	5.1	50.6	0.0	5.1	49.3	
Quebec, Canada	2003	373	51.9	25.2	15.6	7.0	0.3	51.9	0.0	0.3	47.7	
	2007	276	61.2	15.2	13.4	6.9	3.3	61.2	0.0	3.3	64.5	

資料來源：TIMSS 2007 Technical Report, p.201

（四）試題與國家之交互作用（item-by-country interaction）

若高成就國家在某一試題表現上，相較於其餘表現正常的國家，有低成就表現或低學習成就國家在該試題卻有高成就表現，稱為試題在國家間的交互作用，如圖4-3-5至4-3-7所示，各國之95%難度信賴區間的公式列出如下：

$$Upper\,Limit = 1 - \frac{e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \times Z_b}}{1 + e^{RDIFF_{ik} + SE_{RDIFF_{ik}} \times Z_b}}$$

$$Lower\,Limit = 1 - \frac{e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \times Z_b}}{1 + e^{RDIFF_{ik} - SE_{RDIFF_{ik}} \times Z_b}}$$

$RDIFF_{ik}$ 為試題 k 在 i 國家的1PL難度，$SE_{RDIFF_{ik}}$ 為試題 k 在 i 國家單參難度的標準誤，而 Z_b 為經過Bonferroni修正的 Z 分配值。

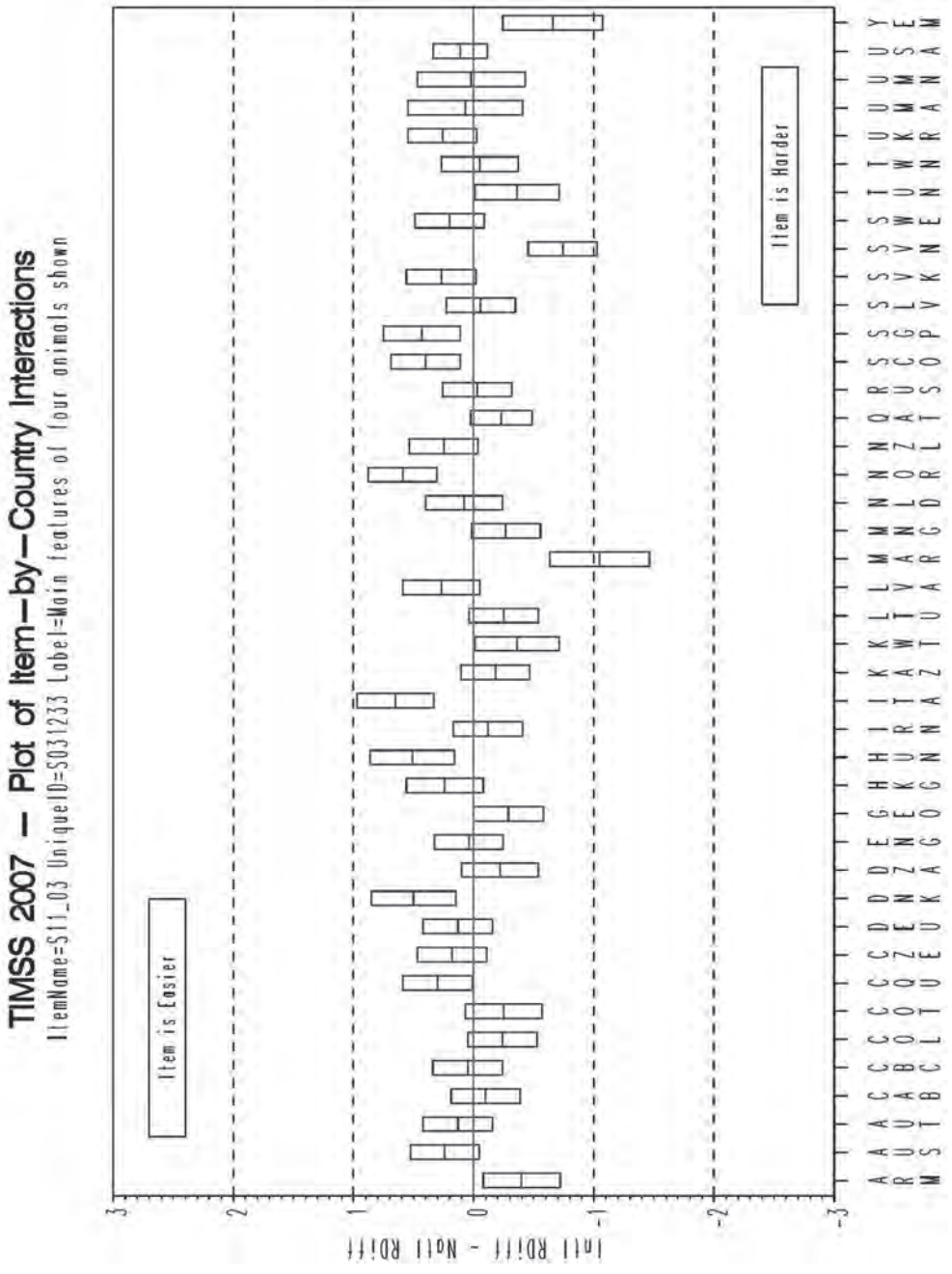


圖4-3-5　試題在各國之難度分布狀況比較

資料來源：TIMSS 2007 Technical Report, p.204

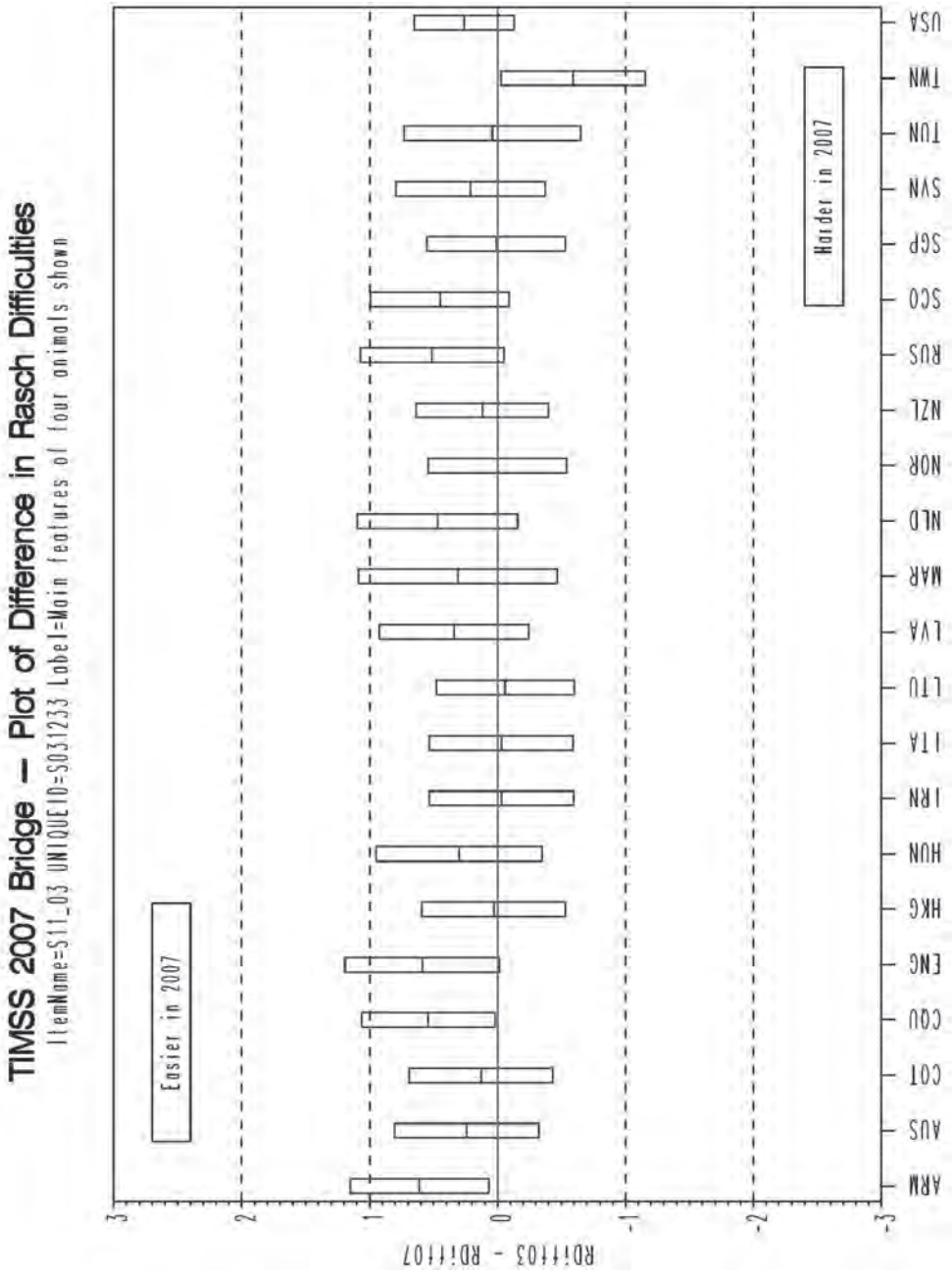


圖4-3-6　定錨題在各國不同年度之難度狀況比較（1）
資料來源：TIMSS 2007 Technical Report, p.206

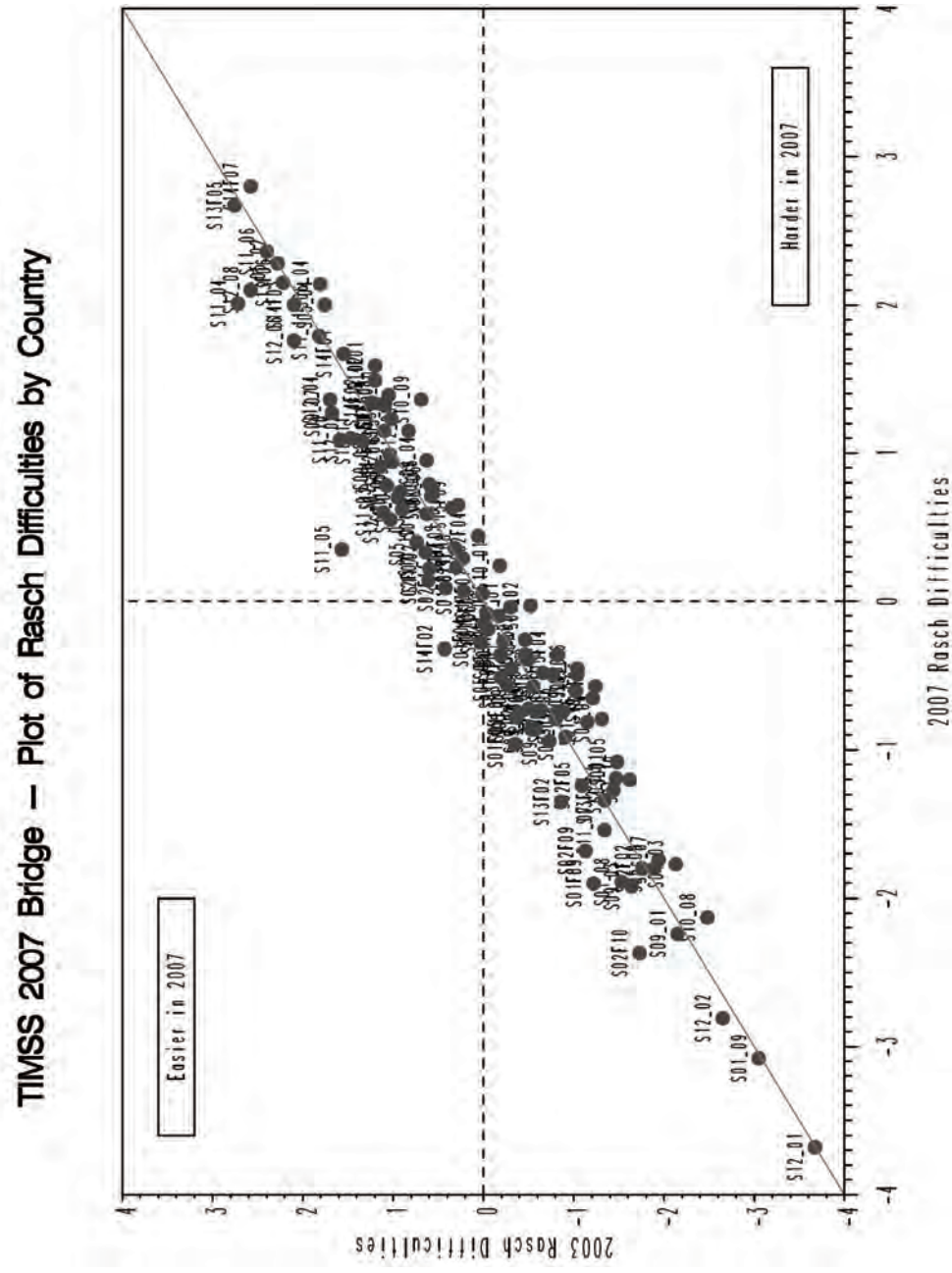


圖4-3-7　定錨題在各國不同年度之難度狀況比較（2）
資料來源：TIMSS 2007 Technical Report, p.207

四、TASA

（一）選擇題

呈現試題之作答人數、答對率、鑑別度、難度、猜測度、點二系列相關、單參難度、選項百分比、難度指標轉換為平均=250、標準差=50的量尺分數等相關數據，如表4-3-5。

（二）開放性試題

呈現試題之作答人數、答對率、選項次數分配、缺失遺漏未作答次數分配等相關數據。

表4-3-5　TASA試題參數估計與描述性統計分析

<table>
<tr><td>科別</td><td>年級</td><td>區塊</td><td colspan="3">題號</td></tr>
<tr><td>數學</td><td>06</td><td>M1</td><td colspan="3">2</td></tr>
<tr><td>題目</td><td colspan="5">下面哪一個比的比值跟其他三個不同？
（1）12：16（2）15：21（3）24：32（4）30：40。</td></tr>
<tr><td>能力指標</td><td colspan="5">略</td></tr>
<tr><td colspan="6">IRT參數估計：鑑別度=☐ 難度=☐ 猜測度=☐ 點二系列相關=☐</td></tr>
<tr><td>答案</td><td colspan="2">2</td><td>認知層次</td><td>略</td><td>編號</td><td>略</td></tr>
<tr><td>古典理論(CTT)選項分析</td><td colspan="5">選　項　1　2*　3　4　其他
選項率　☐ ☐ ☐ ☐ ☐　通過率：☐
高分組　☐ ☐ ☐ ☐ ☐　難　度：☐
低分組　☐ ☐ ☐ ☐ ☐　鑑別度：☐</td></tr>
</table>

資料來源：TASA線上測驗數學科 http://tasa.naer.edu.tw/3online-1-detail.asp?id=3。

五、綜合討論與建議

表4-3-6為TASA與國際性大型資料庫，有關試題一般性統計描述、1PL或3PL參數估計等作一綜合分析比較其中異同。由表4-2-2可以發現，NAEP、TIMSS及TASA對於選擇題的分析所選用的測量模式為三參數對數模式，PISA則為多向度隨機係數多項洛基模式，因此在參數的估計上，或呈現的資料上皆包含了人數、選項百分比、答對率、鑑別度、難度、點二系列相關等數據，有差異且較重要的部份就在PV值的計算應用與缺失遺漏未作答反應的描述。

開放性試題部分，TASA僅呈現人數、無效遺漏未作答百分比或次數分配、選項百分比或次數分配等描述性統計資料，並沒有像NAEP及TIMSS使用二參數對數模式及一般化部份給分模式或PISA所使用的多向度隨機係數多項洛基模式，可見選用適當的測量模式對填充題及開放性試題進行分析為TASA日後資料處理應增加討論的部份。

表4-3-6　試題之描述性統計分析比較

		NAEP	PISA	TIMSS	TASA
選擇題	相同	人數、選項百分比、答對率、鑑別度、難度、點二系列相關			
	相異	遺漏未作答百分比、無效試卷百分比、難度指標轉換為平均=13、標準差=4	試題不合適之均方根的值、遺漏未作答百分比、PV值平均、PV值標準差、期望分數曲線	缺失遺漏未作答百分比	無此資料
開放性試題	相同	人數、無效遺漏未作答百分比或次數分配、選項百分比或次數分配			
	相異	難度指標轉換為平均=13、標準差=4、polyserial相關、Pearson相關	單參難度（Item Map）	鑑別度、單參難度、試題在國家間的交互作用。	無此資料

針對表4-3-6的差異比較可以發現，TASA在PV值的使用、缺失遺漏未作答的資料處理等議題上並無著墨，因此提出底下幾點建議：

1. 增加PV值之計算。
2. 增加缺失、遺漏及未作答選項之描述性統計值。
3. 增加開放性試題難度、鑑別度等參數的估計。
4. 增加不同年度間定錨題之參數比較。

貳、評分者間一致性檢定

一、NAEP

對於開放性試題，在閱讀、寫作、公民測驗中，至少抽取25%的試題；而閱讀、寫作在每個州（郡），至少抽取6%進行兩次評分，表4-3-7為所抽取試題之二次評分一致性狀況比較。

表4-3-7　試題二次評分評分者一致性分析

	Grade	Total Number of Unique Items	Number and Percentage of Items in Percentage Exact Agreement Range							
			60–69%		70–79%		80–89%		Above 90%	
			Number	Percent	Number	Percent	Number	Percent	Number	Percent
Reading										
	4	46	—	—	3	6.5	16	34.7	27	58.6
	8	69	1	1.4	4	5.8	28	40.6	36	52.2
	12	76	1	1.3	4	5.2	36	47.4	35	46.1
Writing										
	4	20	4	20.0	16	80.0	—	—	—	—
	8	23	18	78.3	4	17.4	—	—	—	—
	12	23	10	43.5	9	39.1	3	13.0	—	—
Civics										
	4	21	—	—	3	14.3	11	52.4	7	33.3
	8	28	1	3.6	6	21.4	17	60.7	4	14.3
	12	29	—	—	8	27.6	20	70.0	1	3.4

資料來源：NAEP 1998 Technical Report, p.125

表4-3-8　填充題二次評分評分者一致性分析

Item	Block	Range of Response Codes	Correct Response Codes	Sample Size	Percent Agreement	Cohen's Kappa
R012102	R4	1-2	2	1,923	98	0.970
R012104	R4	1-2	2	1,900	96	0.910
R012106	R4	1-2	2	1,862	93	0.859
R012108	R4	1-2	2	1,761	97	0.920
R012109	R4	1-2	2	1,752	97	0.922
R012112	R4	1-2	2	1,299	94	0.870
R012201	R6	1-2	2	1,925	96	0.923
R012206	R6	1-2	2	1,697	98	0.956
R012208	R6	1-2	2	1,547	93	0.852
R012210	R6	1-2	2	1,452	95	0.820
R012503	R10	1-2	2	1,921	96	0.922
R012504	R10	1-2	2	1,897	98	0.969
R012506	R10	1-2	2	1,865	97	0.949
R012508	R10	1-2	2	1,794	98	0.956
R012511	R10	1-2	2	1,637	97	0.941
R012601	R5	1-2	2	1,897	90	0.759
R012604	R5	1-2	2	1,855	93	0.834
R012611	R5	1-2	2	1,475	89	0.779
R012702	R7	1-2	2	1,908	97	0.913
R012703	R7	1-2	2	1,878	94	0.877
R012705	R7	1-2	2	1,798	94	0.860
R012706	R7	1-2	2	1,765	87	0.705
R012710	R7	1-2	2	1,227	92	0.839
R015802	R9	1-2	2	1,909	92	0.786
R017001	R3	1-2	2	2,035	96	0.902
R017004	R3	1-2	2	1,988	97	0.927
R017006	R3	1-2	2	1,938	96	0.908

資料來源：NAEP 1998 Technical Report, p.565

表4-3-9　多點計分試題二次評分評分者間一致性分析

Item	Block	Range of Response Codes[1]	Sample Size	Percent Agreement	Intraclass Correlation
R012111	R4	1-4	1,555	91	0.946
R012204	R6	1-4	1,894	81	0.906
R012512	R10	1-4	1,591	90	0.957
R012607	R5	1-4	1,770	85	0.867
R012708	R7	1-4	1,637	87	0.908
R015702	R8	1-3	2,036	87	0.841
R015703	R8	1-3	2,017	89	0.862
R015704	R8	1-3	1,978	84	0.870
R015705	R8	1-3	1,963	90	0.942
R015707	R8	1-4	1,834	89	0.904
R015709	R8	1-3	1,558	88	0.881
R015803	R9	1-3	1,891	88	0.841
R015804	R9	1-4	1,845	83	0.873
R015806	R9	1-3	1,706	87	0.884
R015807	R9	1-3	1,548	87	0.880
R015809	R9	1-3	1,389	89	0.858
R017003	R3	1-3	2,019	90	0.917
R017007	R3	1-4	1,868	78	0.899
R017009	R3	1-3	1,613	87	0.821

資料來源：NAEP 1998 Technical Report, p.568

二、TIMSS

針對開放性試題，為了提高測驗成績的信度，在每個國家的測驗題本中，選取約25%的試題進行兩次評分，並在每個國家當中，計算每個試題被評分兩次的次數及前後兩次評分之一致性。

（一）各國國內二次評分評分者一致性分析（within-country scoring reliability）

TIMSS2007計14個題本，每個試題至少出現在兩個題本中，每個題本抽100個學生做二次評分，每個試題進行評分者間一致性檢定的樣本數為200，共1400個學生接受二次評分，約佔總受試者的25%。

表4-3-10　各國國內二次評分評分者間一致性分析

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Algeria	92	58	99	85	54	98
Armenia	99	94	100	97	91	100
Australia	100	98	100	99	95	100
Austria	99	95	100	99	94	100
Chinese Taipei	98	84	100	97	83	100
Colombia	99	93	100	97	89	100
Czech Republic	98	90	100	96	77	100
Denmark	97	83	100	93	74	99
El Salvador	99	96	100	98	85	100
England	99	91	100	98	89	100
Georgia	97	88	100	94	68	100
Germany	97	75	100	95	71	100
Hong Kong SAR	100	98	100	100	98	100
Hungary	100	97	100	99	95	100
Iran, Islamic Rep. of	99	96	100	96	84	100
Italy	99	94	100	99	79	100
Japan	99	94	100	98	84	100
Kazakhstan	99	96	100	99	94	100
Kuwait	100	98	100	98	95	100
Latvia	95	41	100	92	39	100
Lithuania	98	88	100	97	50	100
Morocco	95	33	100	88	29	98
Netherlands	97	86	100	95	72	100
New Zealand	99	95	100	97	90	100
Norway	99	92	100	97	88	100
Qatar	99	91	100	95	78	100
Russian Federation	100	98	100	99	96	100
Scotland	99	91	100	98	87	100
Singapore	99	93	100	97	90	100
Slovak Republic	99	92	100	98	90	100
Slovenia	100	99	100	99	94	100
Sweden	98	89	100	97	87	100
Tunisia	98	86	100	93	77	99
Ukraine	100	98	100	100	98	100
United States	98	83	100	96	72	100
Yemen	98	83	100	93	80	99
International Avg.	98	88	100	96	81	100
Benchmark Participants						
Alberta, Canada	99	93	100	98	90	100
British Columbia, Canada	99	96	100	99	91	100
Dubai, UAE	97	87	100	94	78	100
Massachusetts, US	98	82	100	96	72	100
Minnesota, US	98	79	100	96	68	100
Ontario, Canada	99	88	100	98	88	100
Quebec, Canada	98	90	100	97	86	100

資料來源：TIMSS 2007 Technical Report, p.212

（二）定錨題在各國國內二次評分評分者間一致性分析（trend item scoring reliability）

表4-3-11　定錨題在各國國內二次評分評分者間一致性分析

Countries	Correctness Score Agreement			Diagnostic Score Agreement		
	Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement		Average of Exact Percent Agreement Across Items	Range of Exact Percent Agreement	
		Min	Max		Min	Max
Armenia	96	90	100	93	65	99
Australia	97	84	100	96	83	100
Chinese Taipei	97	93	100	96	88	100
England	98	92	100	97	87	100
Hong Kong SAR	99	93	100	98	87	100
Hungary	99	96	100	97	92	100
Iran, Islamic Rep. of	98	95	100	96	86	100
Japan	98	93	100	96	88	100
Lithuania	97	88	100	94	74	100
Netherlands	97	90	99	95	88	99
New Zealand	98	95	100	97	90	100
Norway	98	96	100	97	93	100
Russian Federation	99	95	100	98	92	100
Scotland	96	91	100	95	90	100
Singapore	95	86	100	93	83	100
Slovenia	96	68	99	93	47	99
Tunisia	98	97	100	95	81	100
United States	98	92	100	96	88	100
International Avg.	97	91	100	96	83	100
Benchmark Participants						
Alberta, Canada	98	91	99	96	85	99
British Columbia, Canada	98	91	99	96	85	99
Ontario, Canada	98	91	99	96	85	99
Quebec, Canada	98	91	99	96	85	99

資料來源：TIMSS 2007 Technical Report, p.216

（三）國際間二次評分評分者間一致性分析（cross-country scoring reliability study）

TIMSS2003為了檢測不同國家的評分者在做二次評分時是否有所差異，從南半球四個國家抽150個學生，每位學生測驗20題數學與21題自然科學；共有6150個作答反應。從北半球是二十個國家中找37位評分者，二次評分有666個組合。每個組合都評分150個學生，共有99900種評分反應。

表4-3-12　國際間二次評分評分者間一致性分析（TIMSS 2007小四數學）

Item Label	Total Valid Comparisons	Exact Percent Agreement	
		Correctness Score Agreement	Diagnostic Score Agreement
M04_02 - M041056	265200	98	96
M04_04 - M041076	265200	99	98
M04_07 - M041146	265200	92	92
M04_09 - M041258A	265200	96	94
M04_09 - M041258B	265200	86	74
M04_11 - M041275	265200	85	85
M05_02 - M031309	265200	99	99
M05_04 - M031242A	265200	98	97
M05_04 - M031242B	265200	97	96
M05_05 - M031247	265200	94	91
M11_02 - M031009	265200	100	99
M11_04 - M031316	265200	99	99
M11_06 - M031079B	261579	99	99
M11_06 - M031079C	261579	97	97
M11_09 - M031325	265200	97	92
M12_04 - M041059	265200	99	95
M12_13 - M041276A	265200	98	98
M12_13 - M041276B	265200	83	79
Average Percent Agreement		95	93

資料來源：TIMSS 2007 Technical Report, p.221

三、綜合討論與建議

表4-3-13為NAEP、TIMSS對開放性試題進行二次評分之評分者間一致性檢定的比較，NAEP及TIMSS分別針對國內及國際間部分開放性試題二次評分資料進行評分者間一致性檢定，藉以提高測驗評量數據之信度，分析方法僅抽樣有部分差異，檢測則皆以一致性百分比為依據，輔以Kappa一致性係數。TASA在國語文、英語文及數學某些年度或年段中亦有多元計分試題之測驗，因此TASA往後對於多元計分試題可增加評分者間一致性檢定，以提高學生學習成就描述之信度。

表4-3-13　評分者間一致性分析比較

NAEP	TIMSS	TASA
對於開放性試題，在閱讀、寫作、公民測驗中，至少抽取25%的試題；而閱讀、寫作在各州（郡）至少抽取6%的試題，進行二次評分的評分者一致性分析。	針對開放性試題在各國測驗題本中，選取約25%的試題進行各國國內、定錨題在各國國內及國際間二次評分的評分者一致性分析。	無此資料

參、試題標記與刪題標準

一、PISA

（一）不良試題

1. 假如某一試題的特徵經過10個以上國家的分析都是不良的，則此試題會被刪除，此種試題又被稱為”不良試題”。
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，但卻顯示試題和國家具有交互作用，即所謂具有差異性的試題，或試題難度對於不同國家而言是不同的。

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。

（二）第二類和第三類不良試題標記

PISA2006針對所謂的不良試題進行標記，藉以區辨出試題不良的原因，以便進行修改、訂正或刪除的作業，如表4-3-14所示。

表4-3-14　不良試題在國家間之標記

Example of summary of dodgy items for a country In Report 3a

PISA 2006 Main Study, Report 3a: Science dodgy items

	Item by country interactions			Discrimination				PISA 2003 link items	
	Number of valid responses	Easier than expected	Harder than expected	Non-key PB is positive	Key PB is negative	Low discrimination	Ability not ordered	Link items	Requires checking
S456Q02	1 437	☐	☐	☐	☐	☑	☐	☐	☐
S476Q01	1 482	☑	☐	☐	☐	☐	☐	☐	☐
S477Q04	1 442	☑	☐	☐	☐	☐	☐	☐	☐
S478Q01	1 443	☐	☐	☑	☐	☐	☐	☐	☐
S493Q01	1 452	☑	☐	☐	☐	☐	☐	☐	☐
S495Q01	1 442	☑	☐	☐	☐	☐	☐	☐	☐
S495Q02	1 440	☐	☑	☐	☐	☐	☐	☐	☐
S508Q02	1 435	☐	☑	☐	☐	☐	☐	☐	☐
S510Q04	1 459	☑	☐	☐	☐	☐	☐	☐	☐
S519Q01	1 438	☐	☑	☐	☐	☐	☐	☐	☐
S524Q06	1 427	☐	☑	☐	☐	☑	☐	☐	☐

資料來源：PISA 2006 Technical Report, p.152

二、TIMSS

（一）使用各種標記藉以突顯試題的特性，底下為整理TIMSS針對試題進行標記的各種充分條件：

1. 難度>95%。
2. 難度<25%（4選1試題）。
3. 誘答力<10%。
4. 錯誤選項的誘答力>正確選項的答對率或錯誤選項的點二系列相關>0。
5. 鑑別度<0.2。
6. 開放性試題鑑別度無遞增。
7. 針對開放性試題，1PL難度>總平均。
8. 針對開放性試題，1PL難度<總平均。
9. 男性與女性在試題難度上有差異。
10. 開放性試題兩次評分的一致性<70%。

（二）符合底下刪題標準之不良試題予以刪除，藉以提高測驗的信度。

1. 已知試題有翻譯上的錯誤，在測試前無及時修正之試題。
2. 選擇題在不同國家間無統一答案。
3. 選擇題正確選項的相關呈現負值或開放性試題鑑別度無遞增。
4. 試題在高（低）成就國家有低（高）表現。
5. 開放性試題，二次評分的一致性（信度）＜70%。
6. 試題在1999年與2003年的測驗表現上產生差異或試題在1999年在某特殊國家已刪除。

三、TASA

（一）不良試題檢測

1. 依據古典測驗理論（classical test theory, CTT）訂定標準如下：

(1) 試題通過率低於0.25，$P_{CTT} < 25\%$。
(2) 試題鑑別度指數介於0 ~ 0.2之間，0<C.T.T.的鑑別度＜0.2。

試題鑑別度指數使用點二系列相關，公式如下：

$$r_j = \frac{(\mu_j - \mu_x)}{\sigma_x}\sqrt{\frac{P_j}{1-P_j}}$$

其中，r_j 指第 j 題鑑別度
μ_j 指答對第 j 題之受試者之測驗總分的平均數
μ_x 指所有受試者之測驗總分的平均數
σ_x 指所有受試者之測驗總分的標準差

2. 依據試題反應理論（item response theory, IRT）訂定標準如下：

(1) 試題鑑別度參數介於0 ~ 0.4之間，$0 < a < 0.4$。

(2) 試題難度參數小於或等於-3，$b \leq -3$ 。

(3) 試題難度參數大於或等於3，$b \geq 3$。

(4) 試題猜測度參數大於或等於0.3，$c \geq 0.3$ 。

（二）符合底下刪題標準之不良試題予以刪除

1. 估計後的試題參數無法收斂。

2. 試題鑑別度指數小於0，C.T.T.的鑑別度＜0。

3. 點二系列相關＜0。

TASA使用的軟體為BILOG-MG（Zimowski, Muraki, Mislevy, & Book, 2003）與SCORIGHT 3.0（Wang, Bradlow, & Wainer, 2004）。其中，刪題標準（1）是指使用上述兩軟體估計試題參數後，試題參數出現異常之情形（例如$a = 8.88$、$b = 13.27$等）；刪題標準（2）是指該試題估計之試題鑑別度指數小於0（$r_j < 0$）。

四、綜合討論與建議

表4-3-15為綜合分析比較TASA與各大型測驗資料庫之間對於不良試題分析上的差異，可以發現因為試題參數估計原則上皆採用1PL或3PL，因此區辨是否為不良試題之參考標準相近，增加對不良試題各項參數的描述或標記，如表4-3-14，使讀者更清楚了解其產生不良效應背後的意涵，或許是TASA往後進行不良試題分析可以著墨之處。

表4-3-15　試題標記與刪題標準比較

	PISA	TIMSS	TASA
不良試題標記	1. 試題在國際間比較過易或過難 2. 錯誤選項的點二系列相關>0 3. 正確選項的點二系列相關<0 4. 鑑別度過低 5. C-R試題鑑別度無遞增	1. P>95% 2. P<25%（4選1試題） 3. 誘答力<10% 4. 錯誤選項的誘答力>正確選項的答對率或錯誤選項的點二系列相關>0 5. D<0.2 6. C-R試題鑑別度無遞增 7. 1-PL難度>總平均 8. 1-PL難度<總平均 9. 性別DIF 10. C-R試題兩次評分的一致性<70%	1. 依據古典測驗理論訂定標準如下： P＜25%、鑑別度指數D介於0~0.2。 2.依據試題反應理論訂定標準如下： 0＜a＜0.4、b≦-3、b≧3、c≧0.3
刪題標準	某一試題的特徵經過10個以上國家的分析都是不良的，則此試題會被刪除	1. 有翻譯上錯誤，測試前無及時修正 2. M-C試題在不同國家無統一答案 3. M-C試題正確選項的相關呈現負值或C-R試題鑑別度無遞增 4. 試題在高（低）成就國家有低（高）表現 5. C-R試題二次評分一致性＜70% 6. 試題在1999年與2003年的測驗表現上產生差異或試題在1999年某特殊國家已刪除	1. 估計後的試題參數無法收斂 2. 試題鑑別度指數C.T.T.的鑑別度＜0 3. 點二系列相關＜0

肆、差異試題功能分析（differential item functioning, DIF）

一、NAEP

對照組及實驗組的選取分三類：男性/女性、白人/黑人及白人/西班牙裔，所使用的分析方法有三種：

（一）The Mantel-Haenszel Procedure

針對選擇題和開放性試題使用不同的分析流程，選擇題使用Mantel-Haenszel chi-square 分析流程（Mantel & Haenszel, 1959）。開放性試題則使用Mantel-Haenszel ordinal 分析流程（Mantel, 1963），它的分析流程乃是比較兩群體間各得分群人數百分比是否存在差異而得。

（二）SIBTEST Procedure

NAEP1998首次針對所有試題同時進行Mantel-Haenszel（Mantel & Haenszel, 1959）及SIBTEST（Shealy & Stout, 1993）差異試題功能分析，如同M-H法，SIBTEST也是利用兩群體間的答對率或選項作答率進行分析，另外再考慮取樣權重的影響，計算標準平均差異（SMD）及效果量係數（SMD/std. dev.），若試題滿足以下兩個條件，則被評為具嚴重試題差異之試題：

1. SMD顯著的偏離0。
2. 效果量係數（SMD/std. dev.）的絕對值大於0.25。

結果發現對於是否存在試題差異的檢測結果一致，M-H法認為存在試題差異之試題，SIBTEST 法依然會將該試題判定存在試題差異，並沒有任何存在嚴重試題差異之試題單獨被SIBTEST 法所判定。

（三）Standardization Method

針對寫作試題，若單獨使用Mantel-Haenszel 法或SIBTEST 法進行試題差異功能分析，所獲取的資訊稍嫌不足，因此NAEP結合M-H法及SIBTEST法中部份分析方法，合併使用稱為Standard NAEP DIF Method （Dorans and Kulick, 1986）。在分析上，寫作試題被視為多點計分，前半段分析流程與M-H法一致，接著計算各得分群兩兩之標準平均差異（SMD），將獲得的所有SMD取平均再除以標準偏差（std. dev.）得到效果量係數（SMD/std. dev.），效果量係數的絕對值大於0.1即被標記出，如表4-3-16所示。

表4-3-16　兩族群間的試題差異功能分析

Items With Absolute SMD (Standardized Mean DIF) > .10

Group	Grade	SMD	ID
NonAcc/Acc	4	-.106	W005402
B/W	4	-.108	W005302
B/W	12	-.129	W009802
B/W	12	.127	W010402
H/W	4	-.101	W004602
H/W	12	-.112	W009202

LEGEND

NonAcc/Acc	Nonaccommodated versus accommodated students
B/W	Black versus White students
H/W	Hispanic versus White students

資料來源：NAEP 1998 Technical Report, p.369

二、綜合討論與建議

表4-3-17是大型測驗之差異試題功能分析比較，相較NAEP及PISA，TASA問卷調查亦包含男女性別及國籍別資料，可評估是否進行DIF檢測。建議TASA往後可針對所有試題，增加性別、城鄉及語言族群間之DIF分析，提高試題品質的控管。

表4-3-17　差異試題功能分析比較

NAEP	PISA	TASA
對照組及實驗組的選取分三類：男性/女性、白人/黑人及白人/西班牙裔，所使用的分析方法有底下三種：The Mantel-Haenszel Procedure、SIBTEST Procedure、Standardization Method	性別、國家間的DIF檢測	無此資料

第四節　問卷背景變項分析

國際性大型測驗資料庫在進行國際間測驗評比時，除了關心學生學習成就表現的變化之外，也會針對學生、學校及教師的背景狀況設計問卷進行調查，藉以分析比較不同的背景變項對學生學習成就表現的影響。TASA在這方面也設計了適合的問卷進行調查，底下是針對問卷類型、內容及分析方法，本節探討內容重點在綜合討論比較NAEP、PISA、TIMSS三個國際大型測驗評比資料庫與TASA在問卷背景變項分析上之差異，並提出具體建議，期望找出一個標準化且合宜的問卷設計和分析流程，以供TASA後續研究分析使用。

壹、問卷類型

表4-4-1　問卷類型比較

NAEP	PISA	TIMSS	TASA
學生問卷 學校問卷 教師問卷	學生問卷 學校問卷	學生問卷 學校問卷 教師問卷	學生問卷 學校問卷

貳、問卷內容

表4-4-2　問卷指標題數比較

問卷	指標	PISA	TIMSS	TASA
學生問卷	性別	1	1	1
	家庭背景	13	5	9
	家庭資源	23	5	10
	做作業時間	1	1	1
	補習家教	7	1	1
	親子關係	0	0	4
	同儕關係	0	5	4
	師生關係	0	5	4
	班級常規	5	0	4
	學習策略	14	0	10
	學習偏好	10	0	9
	學科喜愛度	5	3	2
	學習認知	4	0	2
	學習自信心	6	4	1
	課後活動	0	8	7
	喜愛學習程度	5	0	1
	學習對將來（就業）的影響	5	0	0
	未來是否考慮從事與目前學習相關的行業	4	0	0
學校問卷	校長背景	0	2	3
	教學現況	0	9	1
	學校資源	7	19	7
	師資專長	1	11	1
	學生流動概況	0	0	3
	組織氣氛	0	0	5
	家長與學校互動情形	0	1	6
	學生偏差行為	0	13	8
	教學時數	0	0	1
	新移民教育政策	0	0	10
	其餘教育政策	0	0	6
	經濟弱勢學生比例	0	1	0

	性別	–	1	–
	教師背景	–	8	–
	師資專長	–	18	–
教師問卷	教學時數	–	6	–
	專業發展	–	10	–
	教學態度	–	7	–
	組織氣氛	–	8	–

參、背景變項統計分析

一、問卷背景變項量尺化程序

（一）PISA

使用IRT估算二元計分或多點計分（Likert-type item）試題的潛在特性。

1. 二元計分（IPL）

$$P_i(\theta) = \frac{\exp(\theta_n - \delta_i)}{1 + \exp(\theta_n - \delta_i)}$$

$P_i(\theta)$：學生n在第i題答對的機率；θ_n：學生n的能力值；δ_i：試題i的難度值

2. 多點計分（閾值）

$$P_{x_i}(\theta) = \frac{\exp\left(\sum_{j=0}^{x} \theta_n - \delta_i + \tau_{ij}\right)}{1 + \exp\left(\sum_{j=1}^{k} \theta_n - \delta_i + \tau_{ij}\right)} \quad x_i = 0,1,...,m_i$$

$P_i(\theta)$：學生n在第i題答對的機率；θ_n：學生n的能力值；δ_i：試題i的難度值；τ_{ij}：試題閾值

Figure 16.1
Summed category probabilites for fictitious item

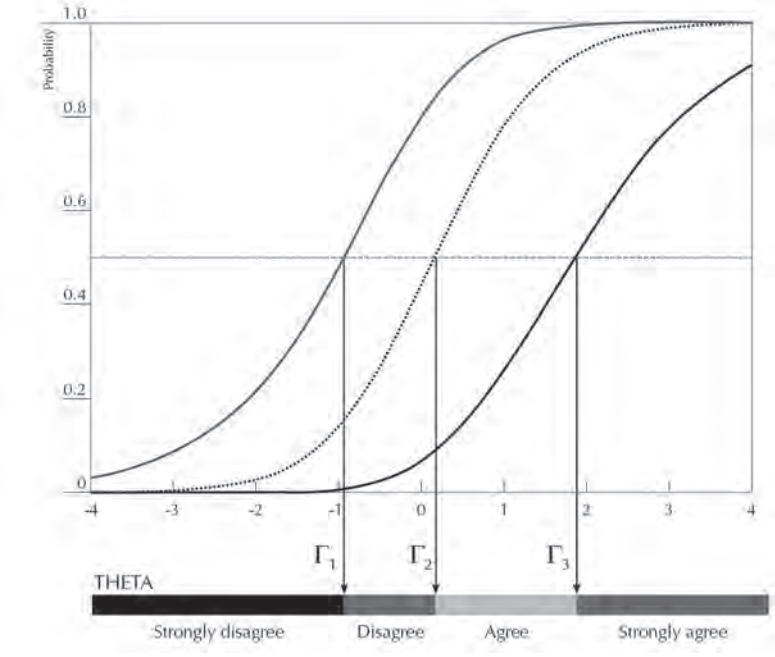


圖4-4-1　試題閾值描述
資料來源：PISA 2006 Technical Report, p.314

3. 國際間試題參數比較

從30個受測國家，每個國家隨機抽500個學生。以及28個受測國家（扣除盧森堡、法國），每個國家隨機抽100個學校。使用WLEs(weighted likelihood estimations)法估計個別學生或學校的能力值。

$$\sum_{i\in\Omega}\left[\left(r_x+\frac{J_n}{2I_n}\right)-\sum_{j=1}^{k}\frac{\exp\left(\sum_{j=0}^{x}\theta_n-\delta_i+\tau_{ij}\right)}{1+\exp\left(\sum_{j=1}^{k}\theta_n-\delta_i+\tau_{ij}\right)}\right]=0$$

4. 標準化θ_n

$$\theta_n{}' = \frac{\theta_n - \bar{\theta}_{OECD}}{\sigma_{\theta(OECD)}}$$

以學生為例，每個國家有θ_1~θ_{500}，全部15000個學生產生一個θ的總平均及標準差，然後再去求$\theta_n{}'$。

（二）綜合討論與建議

針對問卷背景變項分析進行類似試題分析之量尺化程序只有在PISA技術報告中有發現，TIMSS 2007技術報告中並無相關的描述，且PISA測驗模式為單參模式，與TASA現行之三參模式有所不同，因此TASA資料在問卷背景資料是否採用此一分析流程有待商榷。

二、描述性統計量

（一）PISA

表4-4-3為PISA 2006國際成果報告中有關描述性統計分析之數據呈現方式，表中呈現內容為各選項人數百分比與標準差比較。

表4-4-3　選項人數百分比與標準誤（PISA 2006學生問卷）

OECD	Percentages of students reporting that doing well in the subject area is important or very important					
	Science		Reading		Mathematics	
	%	S.E.	%	S.E.	%	S.E.
Australia	72.0	(0.6)	94.5	(0.2)	93.7	(0.3)
Austria	65.4	(1.2)	88.5	(0.8)	91.2	(0.5)
Belgium	64.4	(0.8)	82.2	(0.6)	90.0	(0.6)
Canada	83.4	(0.5)	90.2	(0.3)	95.3	(0.3)
Czech Republic	53.7	(1.3)	88.7	(0.7)	89.3	(0.7)
Denmark	69.5	(0.7)	96.4	(0.3)	96.8	(0.3)
Finland	61.8	(0.8)	79.4	(0.7)	86.0	(0.6)
France	64.1	(1.1)	83.4	(0.7)	89.7	(0.5)
Germany	75.8	(0.8)	92.2	(0.5)	94.5	(0.3)
Greece	74.1	(0.7)	79.9	(0.8)	86.2	(0.6)
Hungary	65.9	(0.9)	82.6	(0.7)	83.4	(0.7)
Iceland	68.0	(0.8)	91.7	(0.4)	97.8	(0.2)
Ireland	74.8	(0.9)	92.9	(0.5)	95.8	(0.4)
Italy	81.9	(0.6)	92.8	(0.3)	90.5	(0.4)
Japan	68.0	(0.8)	88.0	(0.6)	87.2	(0.6)
Korea	75.2	(0.7)	92.4	(0.5)	87.8	(0.6)
Luxembourg	66.8	(0.7)	86.1	(0.5)	84.8	(0.6)
Mexico	88.8	(0.4)	96.3	(0.3)	97.4	(0.3)
Netherlands	72.5	(0.9)	86.5	(0.6)	89.5	(0.5)
New Zealand	75.6	(0.8)	93.2	(0.4)	95.1	(0.3)
Norway	77.4	(0.8)	83.8	(0.7)	91.0	(0.4)
Poland	77.1	(0.7)	88.4	(0.5)	86.4	(0.6)
Portugal	83.0	(0.7)	87.9	(0.6)	89.9	(0.6)
Slovak Republic	60.5	(1.4)	91.1	(0.5)	87.7	(0.8)
Spain	73.6	(0.6)	84.4	(0.4)	88.8	(0.5)
Sweden	72.9	(0.8)	94.1	(0.4)	94.9	(0.4)
Switzerland	62.0	(0.9)	90.1	(0.5)	92.3	(0.5)
Turkey	80.7	(0.9)	93.6	(0.6)	93.0	(0.5)
United Kingdom	83.6	(0.6)	95.3	(0.3)	96.1	(0.2)
United States	82.3	(0.6)	89.7	(0.5)	93.9	(0.4)
OECD average	72.5	(0.2)	89.2	(0.1)	91.2	(0.1)
Chinese Taipei	77.6	(0.7)	87.9	(0.4)	83.4	(0.6)

資料來源：

http://www.pisa.oecd.org/document/2/0,3343,en_32252351_32236191_39718850_1_1_1_1,00.html

（二）TIMSS

表4-4-4為TIMSS2007國際數學報告中有關描述性統計分析之數據呈現方式，表中呈現內容為各選項人數百分比與平均量尺分數表現比較。

表4-4-4　父親教育程度選項人數百分比、平均量尺分數表現及標準誤比較
（TIMSS 2007數學科小四）

Country	University Degree**		Completed Post-secondary Education but Not University		Completed Upper-secondary School	
	Percent of Students	Average Achievement	Percent of Students	Average Achievement	Percent of Students	Average Achievement
Algeria	15 (0.9)	391 (3.6)	12 (0.6)	395 (3.7)	22 (0.7)	391 (2.9)
Armenia	52 (1.7)	505 (4.0)	23 (0.9)	499 (3.8)	16 (1.0)	483 (5.7)
Australia	19 (1.1)	546 (6.1)	23 (0.9)	503 (5.4)	16 (0.7)	484 (3.7)
Bahrain	21 (0.6)	429 (3.1)	8 (0.5)	415 (6.6)	33 (0.9)	402 (2.9)
Bosnia and Herzegovina	15 (1.0)	494 (4.4)	16 (0.6)	471 (4.2)	54 (1.1)	455 (2.6)
Botswana	15 (0.6)	381 (4.1)	17 (0.8)	355 (4.0)	17 (0.7)	358 (3.9)
Bulgaria	29 (1.4)	509 (6.7)	30 (1.3)	469 (4.5)	24 (1.2)	429 (8.6)
Chinese Taipei	20 (1.4)	647 (5.2)	12 (0.7)	633 (5.2)	42 (1.0)	594 (4.1)
Colombia	20 (1.1)	416 (5.4)	9 (0.6)	409 (6.6)	20 (0.7)	380 (4.6)
Cyprus	30 (0.8)	493 (2.7)	12 (0.5)	488 (3.8)	37 (0.7)	461 (2.5)
Czech Republic	17 (0.9)	547 (3.8)	11 (0.5)	512 (3.9)	57 (0.9)	499 (2.5)
Egypt	15 (0.7)	394 (4.6)	19 (1.0)	432 (5.4)	14 (0.6)	408 (6.0)
El Salvador	13 (1.1)	386 (6.4)	9 (0.7)	365 (5.5)	19 (0.9)	350 (3.6)
England	– –	– –	– –	– –	– –	– –
Georgia	47 (2.1)	429 (5.7)	0 (0.0)	~ ~	33 (2.1)	405 (7.0)
Ghana	11 (0.8)	341 (9.9)	20 (0.9)	321 (5.8)	24 (0.9)	314 (4.8)
Hong Kong SAR	13 (1.0)	609 (7.8)	12 (0.6)	587 (7.2)	28 (0.8)	575 (5.6)
Hungary	29 (1.3)	563 (4.6)	13 (0.7)	526 (4.9)	45 (1.2)	505 (3.2)
Indonesia	9 (0.8)	460 (7.7)	6 (0.5)	439 (8.2)	25 (1.2)	412 (5.1)
Iran, Islamic Rep. of	10 (1.0)	469 (9.5)	10 (1.0)	444 (7.7)	18 (1.0)	422 (6.1)
Israel	38 (1.2)	499 (4.3)	10 (0.6)	464 (7.1)	17 (0.8)	441 (6.9)
Italy	21 (1.2)	505 (3.6)	5 (0.4)	491 (6.1)	37 (1.1)	492 (2.8)
Japan	34 (1.0)	606 (3.4)	16 (0.6)	569 (3.7)	27 (1.0)	544 (3.1)
Jordan	29 (1.1)	461 (4.7)	18 (0.9)	455 (4.7)	28 (0.9)	415 (4.5)
Korea, Rep. of	44 (1.4)	627 (3.2)	3 (0.3)	610 (7.1)	39 (1.2)	582 (2.9)
Kuwait	43 (1.4)	370 (3.2)	15 (0.8)	365 (3.8)	26 (0.9)	336 (3.4)
Lebanon	20 (1.3)	490 (5.8)	19 (1.2)	464 (5.6)	16 (1.1)	446 (4.8)
Lithuania	14 (0.8)	549 (4.6)	34 (0.9)	517 (2.9)	23 (1.1)	495 (3.2)
Malaysia	13 (1.0)	510 (7.3)	17 (0.8)	493 (5.4)	34 (0.9)	478 (4.7)
Malta	11 (0.4)	525 (3.6)	11 (0.4)	514 (4.3)	13 (0.5)	513 (3.7)
Norway	39 (1.0)	490 (1.9)	6 (0.4)	469 (5.5)	6 (0.5)	455 (5.8)
Oman	16 (0.9)	388 (5.7)	4 (0.4)	382 (10.6)	18 (0.8)	387 (4.5)
Palestinian Nat'l Auth.	24 (0.9)	398 (5.4)	13 (0.6)	386 (5.8)	35 (0.9)	369 (4.4)
Qatar	48 (0.6)	332 (2.2)	4 (0.2)	310 (8.0)	19 (0.5)	289 (2.7)
Romania	13 (1.0)	524 (5.8)	14 (0.9)	493 (5.8)	44 (1.4)	460 (4.6)
Russian Federation	38 (1.3)	540 (4.4)	34 (1.3)	511 (5.1)	12 (1.0)	471 (6.2)
Saudi Arabia	31 (1.2)	354 (3.6)	5 (0.5)	343 (9.4)	20 (0.9)	325 (3.9)
Scotland	– –	– –	– –	– –	– –	– –
Serbia	20 (1.2)	533 (4.6)	16 (0.8)	496 (5.0)	51 (1.3)	477 (3.8)
Singapore	20 (0.7)	646 (3.9)	19 (0.6)	603 (4.7)	28 (0.7)	587 (4.3)
Slovenia	24 (0.9)	532 (3.3)	35 (1.0)	503 (2.6)	15 (0.7)	486 (4.6)
Sweden	19 (0.8)	515 (3.3)	13 (0.6)	510 (3.2)	13 (0.6)	487 (3.4)
Syrian Arab Republic	15 (0.9)	419 (4.7)	22 (0.9)	414 (4.8)	23 (0.8)	385 (4.3)
Thailand	12 (1.1)	522 (11.7)	5 (0.3)	481 (9.7)	14 (0.6)	455 (6.5)
Tunisia	13 (1.1)	459 (4.6)	17 (0.9)	437 (3.9)	25 (1.0)	414 (2.9)
Turkey	7 (0.8)	558 (8.7)	3 (0.3)	497 (8.8)	20 (1.2)	470 (5.2)
Ukraine	40 (1.4)	494 (4.3)	34 (0.9)	465 (3.6)	12 (0.8)	417 (6.6)
United States	44 (1.2)	531 (3.3)	7 (0.4)	503 (4.1)	21 (0.6)	495 (2.3)
‡ Morocco	20 (1.3)	407 (4.9)	0 (0.0)	~ ~	18 (1.0)	394 (5.8)
International Avg.	24 (0.2)	485 (0.9)	14 (0.1)	467 (1.0)	25 (0.1)	444 (0.9)

資料來源：TIMSS 2007 International Mathematics Report, p.146

表4-4-4　父親教育程度選項人數百分比、平均量尺分數表現及標準誤比較
（TIMSS 2007數學科小四）

Country	Completed Lower-secondary School		Less than Lower-secondary School		Do Not Know	
	Percent of Students	Average Achievement	Percent of Students	Average Achievement	Percent of Students	Average Achievement
Algeria	26 (0.8)	379 (2.0)	19 (1.2)	385 (3.3)	6 (0.3)	386 (4.6)
Armenia	2 (0.4)	~ ~	1 (0.2)	~ ~	6 (0.5)	482 (5.8)
Australia	14 (0.9)	474 (4.5)	1 (0.2)	~ ~	28 (0.9)	487 (5.0)
Bahrain	15 (0.6)	367 (4.2)	6 (0.5)	383 (6.9)	18 (0.6)	388 (3.1)
Bosnia and Herzegovina	12 (0.9)	411 (5.6)	1 (0.3)	~ ~	3 (0.3)	421 (9.0)
Botswana	18 (0.6)	359 (3.5)	14 (0.7)	368 (3.9)	20 (0.8)	381 (3.9)
Bulgaria	8 (1.1)	418 (12.3)	1 (0.2)	~ ~	9 (0.7)	451 (8.1)
Chinese Taipei	14 (0.9)	554 (6.2)	3 (0.4)	543 (11.9)	9 (0.5)	554 (9.9)
Colombia	23 (0.9)	365 (5.0)	23 (1.2)	355 (3.8)	6 (0.5)	365 (7.5)
Cyprus	9 (0.4)	437 (4.6)	4 (0.3)	413 (6.5)	7 (0.6)	418 (6.4)
Czech Republic	2 (0.2)	~ ~	0 (0.0)	~ ~	13 (0.6)	466 (3.7)
Egypt	29 (1.1)	381 (4.6)	14 (0.8)	363 (6.3)	10 (0.7)	370 (6.2)
El Salvador	39 (1.3)	326 (3.4)	16 (1.1)	323 (3.5)	4 (0.4)	323 (7.9)
England	– –	– –	– –	– –	– –	– –
Georgia	2 (0.3)	~ ~	0 (0.1)	~ ~	18 (1.2)	383 (10.6)
Ghana	27 (1.2)	298 (5.1)	12 (0.8)	305 (7.6)	6 (0.6)	297 (8.5)
Hong Kong SAR	29 (0.9)	563 (7.3)	3 (0.3)	567 (11.1)	16 (0.8)	547 (7.6)
Hungary	7 (0.9)	434 (7.7)	1 (0.2)	~ ~	5 (0.6)	499 (7.6)
Indonesia	24 (0.9)	380 (4.2)	28 (1.4)	380 (4.9)	9 (0.6)	369 (6.7)
Iran, Islamic Rep. of	28 (1.0)	392 (4.4)	31 (1.5)	376 (4.3)	3 (0.3)	356 (9.5)
Israel	7 (0.6)	409 (9.5)	3 (0.4)	404 (12.3)	26 (1.0)	458 (5.7)
Italy	24 (1.1)	457 (4.7)	3 (0.3)	420 (9.8)	10 (0.7)	443 (5.6)
Japan	2 (0.2)	~ ~	0 (0.1)	~ ~	21 (0.8)	553 (3.4)
Jordan	9 (0.5)	389 (8.7)	9 (0.8)	390 (8.6)	7 (0.6)	388 (11.4)
Korea, Rep. of	3 (0.3)	548 (9.9)	1 (0.1)	~ ~	10 (0.6)	545 (5.0)
Kuwait	0 (0.0)	~ ~	16 (0.9)	334 (4.3)	0 (0.0)	~ ~
Lebanon	13 (1.0)	425 (5.6)	19 (1.6)	425 (6.0)	13 (0.9)	446 (5.3)
Lithuania	4 (0.5)	436 (6.3)	0 (0.1)	~ ~	24 (1.0)	492 (4.0)
Malaysia	19 (0.9)	454 (4.8)	7 (0.6)	450 (8.5)	11 (1.0)	441 (9.1)
Malta	34 (0.7)	477 (2.2)	3 (0.3)	460 (9.7)	27 (0.6)	470 (3.1)
Norway	2 (0.2)	~ ~	1 (0.1)	~ ~	46 (0.9)	460 (2.3)
Oman	17 (0.7)	381 (4.3)	31 (1.1)	370 (3.4)	14 (0.9)	345 (6.8)
Palestinian Nat'l Auth.	11 (0.6)	347 (5.7)	9 (0.7)	340 (5.7)	8 (0.6)	323 (8.9)
Qatar	13 (0.4)	270 (3.5)	7 (0.3)	284 (3.8)	9 (0.4)	295 (4.1)
Romania	9 (1.0)	424 (8.0)	2 (0.4)	~ ~	17 (1.0)	436 (5.0)
Russian Federation	5 (0.5)	462 (8.7)	0 (0.1)	~ ~	10 (0.8)	487 (6.3)
Saudi Arabia	17 (0.9)	315 (5.0)	23 (1.2)	310 (4.5)	5 (0.5)	335 (7.8)
Scotland	– –	– –	– –	– –	– –	– –
Serbia	7 (0.9)	421 (10.5)	0 (0.1)	~ ~	5 (0.4)	456 (7.6)
Singapore	6 (0.4)	567 (7.8)	6 (0.4)	553 (7.2)	21 (0.7)	564 (6.2)
Slovenia	4 (0.4)	465 (7.7)	1 (0.1)	~ ~	22 (0.9)	497 (2.7)
Sweden	4 (0.3)	473 (5.1)	1 (0.2)	~ ~	50 (1.1)	484 (2.9)
Syrian Arab Republic	25 (1.0)	386 (4.8)	11 (0.8)	384 (7.2)	4 (0.4)	378 (9.7)
Thailand	26 (0.9)	421 (4.6)	26 (1.6)	429 (7.3)	18 (1.1)	417 (4.8)
Tunisia	25 (1.0)	402 (3.3)	12 (0.9)	411 (3.5)	8 (0.5)	423 (4.7)
Turkey	52 (1.3)	412 (4.8)	16 (1.0)	389 (4.7)	1 (0.2)	~ ~
Ukraine	5 (0.4)	401 (7.0)	0 (0.1)	~ ~	8 (0.6)	432 (7.0)
United States	7 (0.5)	467 (4.1)	2 (0.2)	~ ~	18 (0.5)	496 (3.3)
‡ Morocco	16 (1.0)	369 (4.5)	36 (1.7)	368 (3.3)	10 (0.9)	367 (7.9)
International Avg.	15 (0.1)	418 (1.0)	9 (0.1)	396 (1.4)	13 (0.1)	431 (1.1)

資料來源：TIMSS 2007 International Mathematics Report, p.147

（三）綜合討論與建議

國際大型測驗評比資料庫在進行平均量尺分數與問卷背景變項關聯性分析時，所使用的能力值參數為可能值，藉以降低測量誤差，此點亦為TASA進行後續分析研究時首要克服的問題。表4-4-5為TASA與其他國際大型測驗評比資料庫對問卷背景變項所進行的描述性統計分析之差異比較，可以發現，選項人數百分比及平均量尺分數為共同呈現的數據，但是對於量尺分數的取得方式卻有所不同，TASA採用的是三參點估計的能力值，NAEP、PISA及TIMSS所使用的能力值參數則為可能值，藉以計算抽樣誤差及測量誤差，以得到標準誤之數據。

表4-4-5　問卷之描述性統計量比較

NAEP	PISA	TIMSS	TASA
平均量尺分數 標準誤	選項人數百分比 平均量尺分數 標準誤	選項人數百分比 平均量尺分數 標準誤	選項人數百分比 平均量尺分數 標準差

三、問卷題項信、效度分析

表4-4-6為整理應用於基礎研究之問卷有關信度強度與Cronbach's alpha係數參照表。

表4-4-6　信度強度與Cronbach's alpha係數參照表

可信度	Cronbach's α 係數
不可信	Cronbach's α 係數＜0.6
可接受	0.6＜Cronbach's α 係數＜0.8
可信	0.8＜Cronbach's α 係數

資料來源：SPSS統計應用學習實務（頁5-4），吳明隆，2006，臺北市：知城

表4-4-7　社經地位指標之因素負荷量與信度分析（PISA 2006數學）

Factor loadings and internal consistency of ESCS 2006 in OECD countries

	Factor loadings			
	HISEI	PARED	HOMEPOS	Reliability[1]
Australia	0.80	0.78	0.67	0.59
Austria	0.81	0.78	0.71	0.64
Belgium	0.83	0.80	0.71	0.68
Canada	0.79	0.78	0.67	0.60
Czech Republic	0.84	0.78	0.70	0.65
Denmark	0.79	0.78	0.70	0.63
Finland	0.77	0.75	0.63	0.52
France	0.82	0.79	0.73	0.67
Germany	0.81	0.76	0.72	0.64
Greece	0.84	0.82	0.72	0.71
Hungary	0.83	0.85	0.77	0.74
Iceland	0.80	0.80	0.59	0.57
Ireland	0.81	0.79	0.74	0.67
Italy	0.84	0.81	0.73	0.71
Japan	0.72	0.77	0.68	0.53
Korea	0.76	0.81	0.75	0.66
Luxembourg	0.83	0.81	0.73	0.69
Mexico	0.85	0.86	0.82	0.80
Netherlands	0.82	0.78	0.75	0.68
New Zealand	0.79	0.76	0.69	0.59
Norway	0.78	0.77	0.66	0.55
Poland	0.87	0.86	0.74	0.73
Portugal	0.86	0.85	0.80	0.77
Slovakia	0.85	0.82	0.74	0.72
Spain	0.84	0.82	0.70	0.69
Sweden	0.77	0.73	0.70	0.57
Switzerland	0.80	0.78	0.68	0.62
Turkey	0.80	0.83	0.79	0.72
United Kingdom	0.78	0.75	0.71	0.60
United States	0.80	0.81	0.74	0.67
Median	**0.81**	**0.79**	**0.72**	**0.67**

1. Reliabilities (Standardised Cronbach's alpha) computed with weighted national samples.

資料來源：PISA 2006 Technical Report, p.347

（一）PISA

PISA 2006數據呈現信度分析之Cronbach's alpha值，詳如表4-4-7。

表4-4-8　學習自信心指標之信度與效度分析（TIMSS 2007數學）

Countries	Grade 4						Grade 8					
	Cronbach's Alpha Between the Component Variables		Multiple R Between Student Achievement and Component Variables		Percent of Variance in Student Achievement Accounted for by the Component Variables		Cronbach's Alpha Between the Component Variables		Multiple R Between Student Achievement and Component Variables		Percent of Variance in Student Achievement Accounted for by the Component Variables	
	Mathematics	Science	Mathematics	Science	Mathematics	Science	Mathematics	Science	Mathematics	Science	Mathematics	Science
Algeria	0.36	0.41	0.28	0.27	0.08	0.08	0.54	–	0.44	–	0.20	–
Armenia	0.60	0.61	0.17	0.17	0.03	0.03	0.66	–	0.21	–	0.04	–
Australia	0.75	0.74	0.46	0.27	0.22	0.07	0.81	0.81	0.55	0.37	0.30	0.14
Austria	0.78	0.75	0.47	0.34	0.22	0.12	◊	◊	◊	◊	◊	◊
Bahrain	◊	◊	◊	◊	◊	◊	0.67	0.58	0.51	0.46	0.26	0.21
Bosnia and Herzegovina	◊	◊	◊	◊	◊	◊	0.78	–	0.51	–	0.27	–
Botswana	◊	◊	◊	◊	◊	◊	0.46	0.43	0.29	0.36	0.09	0.13
Bulgaria	◊	◊	◊	◊	◊	◊	0.70	–	0.42	–	0.18	–
Chinese Taipei	0.73	0.73	0.47	0.26	0.22	0.07	0.84	0.81	0.55	0.41	0.31	0.17
Colombia	0.43	0.46	0.36	0.35	0.13	0.12	0.68	0.63	0.37	0.30	0.14	0.09
Cyprus	◊	◊	◊	◊	◊	◊	0.79	–	0.52	–	0.28	–
Czech Republic	0.75	0.77	0.43	0.31	0.18	0.10	0.85	–	0.53	–	0.28	–
Denmark	0.78	0.76	0.43	0.26	0.18	0.07	◊	◊	◊	◊	◊	◊
Egypt	◊	◊	◊	◊	◊	◊	0.46	0.53	0.35	0.40	0.12	0.16
El Salvador	0.35	0.33	0.33	0.36	0.11	0.13	0.57	0.37	0.36	0.34	0.13	0.12
England	0.75	0.79	0.44	0.29	0.20	0.08	0.79	0.84	0.46	0.37	0.21	0.14
Georgia	0.51	0.56	0.34	0.25	0.12	0.06	0.66	–	0.38	–	0.14	–
Germany	0.81	0.76	0.49	0.35	0.24	0.13	◊	◊	◊	◊	◊	◊
Ghana	◊	◊	◊	◊	◊	◊	0.51	0.52	0.33	0.35	0.11	0.12
Hong Kong SAR	0.72	0.68	0.40	0.29	0.16	0.09	0.80	0.75	0.38	0.26	0.15	0.07
Hungary	0.78	0.79	0.51	0.39	0.26	0.15	0.84	–	0.56	–	0.31	–
Indonesia	◊	◊	◊	◊	◊	◊	0.43	–	0.30	–	0.09	–
Iran, Islamic Rep. of	0.73	0.78	0.48	0.44	0.23	0.19	0.74	0.73	0.46	0.35	0.21	0.13
Israel	◊	◊	◊	◊	◊	◊	0.73	0.74	0.41	0.44	0.17	0.20
Italy	0.69	0.68	0.35	0.24	0.12	0.06	0.84	0.81	0.48	0.31	0.23	0.10
Japan	0.76	0.75	0.47	0.28	0.22	0.08	0.78	0.79	0.50	0.40	0.25	0.16
Jordan	◊	◊	◊	◊	◊	◊	0.65	0.62	0.52	0.42	0.27	0.18
Kazakhstan	0.79	0.79	0.28	0.21	0.08	0.04	◊	◊	◊	◊	◊	◊
Korea, Rep. of	◊	◊	◊	◊	◊	◊	0.86	0.86	0.64	0.48	0.40	0.23
Kuwait	0.35	0.42	0.38	0.39	0.14	0.15	0.59	0.53	0.43	0.34	0.18	0.11
Latvia	0.72	0.71	0.50	0.32	0.25	0.10	◊	◊	◊	◊	◊	◊
Lebanon	◊	◊	◊	◊	◊	◊	0.65	–	0.46	–	0.21	–
Lithuania	0.71	0.70	0.54	0.34	0.29	0.12	0.79	–	0.58	–	0.33	–
Malaysia	◊	◊	◊	◊	◊	◊	0.64	0.66	0.40	0.28	0.16	0.08
Malta	◊	◊	◊	◊	◊	◊	0.78	–	0.47	–	0.22	–
Morocco	0.44	0.42	0.28	0.28	0.08	0.08	0.63	–	0.45	–	0.20	–
Netherlands	0.82	0.78	0.43	0.29	0.18	0.08	◊	◊	◊	◊	◊	◊
New Zealand	0.69	0.68	0.48	0.35	0.23	0.12	◊	◊	◊	◊	◊	◊
Norway	0.68	0.72	0.39	0.32	0.15	0.10	0.80	0.79	0.61	0.37	0.38	0.13
Oman	◊	◊	◊	◊	◊	◊	0.49	0.49	0.46	0.43	0.21	0.19
Palestinian Nat'l Auth.	◊	◊	◊	◊	◊	◊	0.54	0.53	0.46	0.45	0.21	0.20
Qatar	0.41	0.47	0.36	0.36	0.13	0.13	0.61	0.53	0.40	0.34	0.16	0.12
Romania	◊	◊	◊	◊	◊	◊	0.63	–	0.46	–	0.21	–
Russian Federation	0.74	0.75	0.38	0.28	0.14	0.08	0.84	–	0.52	–	0.27	–
Saudi Arabia	◊	◊	◊	◊	◊	◊	0.49	0.48	0.44	0.44	0.19	0.19
Scotland	0.72	0.74	0.32	0.25	0.10	0.06	0.77	0.83	0.45	0.48	0.20	0.23
Serbia	◊	◊	◊	◊	◊	◊	0.82	–	0.64	–	0.41	–
Singapore	0.76	0.75	0.50	0.31	0.25	0.10	0.82	0.82	0.45	0.26	0.20	0.07
Slovak Republic	0.73	0.73	0.45	0.36	0.21	0.13	◊	◊	◊	◊	◊	◊
Slovenia	0.66	0.65	0.50	0.32	0.25	0.10	0.76	–	0.54	–	0.30	–
Sweden	0.72	0.73	0.38	0.29	0.15	0.09	0.82	–	0.58	–	0.33	–
Syrian Arab Republic	◊	◊	◊	◊	◊	◊	0.57	–	0.42	–	0.17	–
Thailand	◊	◊	◊	◊	◊	◊	0.58	0.61	0.28	0.24	0.08	0.06
Tunisia	0.45	0.49	0.47	0.43	0.22	0.19	0.73	0.62	0.48	0.39	0.23	0.16
Turkey	◊	◊	◊	◊	◊	◊	0.76	0.71	0.50	0.38	0.25	0.14
Ukraine	0.69	0.68	0.46	0.33	0.21	0.11	0.79	–	0.53	–	0.28	–
United States	0.76	0.78	0.46	0.34	0.21	0.12	0.84	0.82	0.46	0.34	0.21	0.12
Yemen	0.09	0.31	0.22	0.23	0.05	0.05	◊	◊	◊	◊	◊	◊
International Median	0.72	0.72	0.43	0.31	0.18	0.10	0.73	0.66	0.46	0.37	0.21	0.14

資料來源：TIMSS 2007 Technical Report, p.292

（二）TIMSS

TIMSS2007呈現各項指標內信度Cronbach's alpha值、效度複相關R值及解釋變異量R^2，詳如表4-4-8。

（三）綜合討論與建議

表4-4-9為比較PISA及TIMSS針對問卷背景變項各項指標進行信度與效度分析之差異，可提供TASA後續研究參考使用。

表4-4-9　問卷之Cronbach's alpha、相關分析及解釋變異量R^2比較

PISA	TIMSS	TASA
信度Cronbach's alpha值	信度Cronbach's alpha值、效度複相關R值及解釋變異量R2	無此資料

四、顯著性檢定

（一）TIMSS

TIMSS 2003在問卷分析中有使用變異數分析與卡方檢定，但是於TIMSS 2007技術報告與國際報告中並未提及使用變異數分析，只陳述針對各指標進行信度、效度及模式適配度檢定，而顯著性檢定則以平均數加減兩倍標準誤代表95%之信賴區間做為分析方法，這或許是因為大型測驗樣本數龐大，進行變異數分析容易增加型I錯誤發生的風險，統計上的意義微弱所致，這有賴更多文獻資料的探討與佐證。

（二）綜合討論與建議

表4-4-10為TASA與其他國際大型測驗評比資料庫針對問卷背景題項各選項百分比或平均量尺分數之顯著性檢定差異比較，可以發現其中有相當大的不同，一般而言，大型資料庫樣本數量龐大，統計上常用的假設檢定容易增加型I錯誤發生的風險，因此NAEP、PISA及TIMSS針對學生能力值皆採用可能值之估計，藉以計算抽樣誤差、測量誤差及標準誤，再藉由平均量尺分數加減兩倍的標準誤代表95%的信賴區間來進行群體間顯著性差異的檢定，如此可以降低統計上偏誤造成的影響。因此建議TASA亦採用標準誤計算之方式進行顯著性檢定，如此降低抽樣誤差及測量誤差造成之影響。

表4-4-10　顯著性檢定比較

NAEP	PISA	TIMSS	TASA
平均數加減兩倍的標準誤代表95%的信賴區間	平均數加減兩倍的標準誤代表95%的信賴區間	平均數加減兩倍的標準誤代表95%的信賴區間	獨立樣本T檢定 單因子變異數分析

表4-4-11　家庭社經地位指標之主成分分析（PISA 2006數學）

	Factor loadings		
	HISEI	PARED	HOMEPOS
Australia	0.80	0.78	0.67
Austria	0.81	0.78	0.71
Belgium	0.83	0.80	0.71
Canada	0.79	0.78	0.67
Czech Republic	0.84	0.78	0.70
Denmark	0.79	0.78	0.70
Finland	0.77	0.75	0.63
France	0.82	0.79	0.73
Germany	0.81	0.76	0.72
Greece	0.84	0.82	0.72
Hungary	0.83	0.85	0.77
Iceland	0.80	0.80	0.59
Ireland	0.81	0.79	0.74
Italy	0.84	0.81	0.73
Japan	0.72	0.77	0.68
Korea	0.76	0.81	0.75
Luxembourg	0.83	0.81	0.73
Mexico	0.85	0.86	0.82
Netherlands	0.82	0.78	0.75
New Zealand	0.79	0.76	0.69
Norway	0.78	0.77	0.66
Poland	0.87	0.86	0.74
Portugal	0.86	0.85	0.80
Slovakia	0.85	0.82	0.74
Spain	0.84	0.82	0.70
Sweden	0.77	0.73	0.70
Switzerland	0.80	0.78	0.68
Turkey	0.80	0.83	0.79
United Kingdom	0.78	0.75	0.71
United States	0.80	0.81	0.74
Median	0.81	0.79	0.72

資料來源：PISA 2006 Technical Report, p.347

五、主成分分析（PCA）

（一）PISA

PISA自2000年起，針對家庭社經地位指標（ESCS）使用主成分分析，以第一主成分軸之因素分數代表學生在家庭社經地位指標的得分。家庭社經地位指標包含父母親職業、教育程度及家中資源等三項。然後對三個項目的得分進行主成份分析，取第一主成份之因素分數定義為社經地位之得分，公式如下：

$$ESCS = \frac{\beta_1 HISEI' + \beta_2 PARED' + \beta_3 HOMEPOS'}{\varepsilon_f}$$

β1、 β2 和 β3 為各國在父母親職業、教育程度及家中資源三項上的因素負荷量，εf 為第一主成分軸之特徵值。表4-4-11為PISA 2006各國社經地位指標主成分分析之因素負荷量整理。

（二）TASA

針對家庭社經地位指標使用主成分分析，以第一主成分軸之因素分數來代表學生在家庭社經地位指標的得分。對於家庭社經地位指標的評定，主要是調查學生家中扶養者（親生父親（或繼父、養父）、親生母親（或繼母、養母））的教育程度與家庭資源等。首先，在教育程度調查方面，分別就教育年數之長短進行分數轉換，計分依續由低至高為0分（小學沒畢業或沒有上過學）、6分（國小畢業）、9分（國中畢業）、12分（高中/職畢業）、14分（專科畢業）、16分（大學畢業）、18分（碩士以上學位）等，分數愈高代表教育程度愈高。其次，在家庭資源調查方面，主要依家中是否有學習相關之設備、參加課輔、才藝班等，給與1分（有）、0分（沒有）之計分，其中，家中的課外讀物方面，依其調查中的0-10本、11-25本、26-100本、101-200本、200本以上等5個選項，以100本為分界點建議，分別將以下者轉換計分為0、以上者計分為1外，此外，並另調查學生是否會和家人一起參加各種文藝活動的頻率，分別給與1至4分不等。最後，將勾選加總計算之總分，作為家庭教育資源之測量指標，分數愈高者，代表家中資源愈高。如此，進而再採用主成份分析，將上述父親學歷、母親學歷與家庭資源等三項資料，投入分析，取第一主成份之因素分數為此所定義之社經地位得分。

（三）綜合討論與建議

由表4-4-12之比較可知，PISA 2006及TIMSS 2007為了進行可能值之計算，針對問卷背景變項皆進行因素縮減的處理，分別以95%及90%的總解釋變異量

為篩選標準，但是TASA問卷試題數遠比此兩大資料庫來的少，或許分析可以納入全部之問卷背景變項題項。另外針對家庭社經地位指標所進行的主成分分析，建議TASA可以再增加信度的分析，使分析內容更形完整。

表4-4-12　問卷之主成分分析比較

NAEP	PISA	TIMSS	TASA
無此資料	針對家庭社經地位指標使用主成分析，以第一主成分軸之因素分數來代表學生在家庭社經地位指標的得分並進行指標信度分析（Cronbach's alpha值）。	無此資料	針對家庭社經地位指標使用主成分析，以第一主成分軸之因素分數來代表學生在家庭社經地位指標的得分，接著再利用社經地位指標得分與學生量尺分數進行Pearson相關分析。

表4-4-13　記憶複述、控制及精緻化學習策略模式適配度檢測與潛在相關比較（PISA 2003數學科小四）

	Model fit				Latent correlations between:		
	RMSEA	RMR	CFI	NNFI	CSTRAT/ ELAB	CSTRAT/ MEMOR	ELAB/ MEMOR
Australia	0.066	0.023	0.90	0.90	0.75	0.96	0.78
Austria	0.055	0.041	0.92	0.92	0.45	0.84	0.35
Belgium	0.071	0.037	0.87	0.87	0.62	0.88	0.60
Canada	0.082	0.042	0.89	0.89	0.60	0.89	0.63
Czech Republic	0.061	0.021	0.91	0.91	0.81	0.94	0.83
Denmark	0.063	0.030	0.92	0.92	0.60	0.95	0.60
Finland	0.074	0.039	0.82	0.83	0.71	0.90	0.68
France	0.068	0.029	0.83	0.83	0.50	0.81	0.49
Germany	0.074	0.024	0.90	0.90	0.84	0.89	0.73
Greece	0.058	0.029	0.90	0.90	0.65	0.95	0.57
Hungary	0.071	0.034	0.86	0.86	0.62	0.85	0.60
Iceland	0.076	0.030	0.89	0.89	0.78	1.03	0.86
Ireland	0.093	0.046	0.85	0.85	0.81	1.00	0.76
Italy	0.057	0.040	0.93	0.93	0.49	0.94	0.50
Japan	0.066	0.022	0.92	0.92	0.89	0.91	0.86
Korea	0.056	0.026	0.92	0.92	0.56	0.96	0.74
Luxembourg	0.074	0.028	0.87	0.87	0.61	0.87	0.68
Mexico	0.077	0.040	0.87	0.87	0.80	0.97	0.86
Netherlands	0.070	0.024	0.87	0.87	0.75	0.93	0.73
New Zealand	0.067	0.028	0.85	0.85	0.62	0.94	0.48
Norway	0.089	0.037	0.88	0.88	0.74	0.78	0.62
Poland	0.066	0.030	0.92	0.92	0.71	1.03	0.75
Portugal	0.073	0.029	0.87	0.87	0.82	0.80	0.79
Slovak Republic	0.067	0.036	0.88	0.88	0.43	0.73	0.51
Spain	0.062	0.039	0.91	0.91	0.43	0.76	0.48
Sweden	0.065	0.028	0.94	0.94	0.92	0.99	0.86
Switzerland	0.071	0.029	0.80	0.81	0.63	0.64	0.18
Turkey	0.061	0.028	0.92	0.92	0.65	0.89	0.62
United Kingdom	0.063	0.030	0.93	0.93	0.75	0.89	0.82
United States	0.075	0.031	0.92	0.92	0.79	0.98	0.87
OECD	0.054	0.023	0.94	0.92	0.66	0.90	0.67

Note: Model estimates based on international student calibration sample (500 students per OECD country).

資料來源：PISA 2003 Technical Report, p.296

六、模式適配度檢定

（一）PISA

使用驗證性因素分析（confirmatory factor analysis）──LISREL軟體，參考指標如下：

1. Root-Mean Square Error of Approximation（RMSEA）

   RMSEA＞0.1表示適配度不佳；0.05＜RMSEA＜0.1表示適配度可接受；RMSEA＜0.05表示適配度良好。

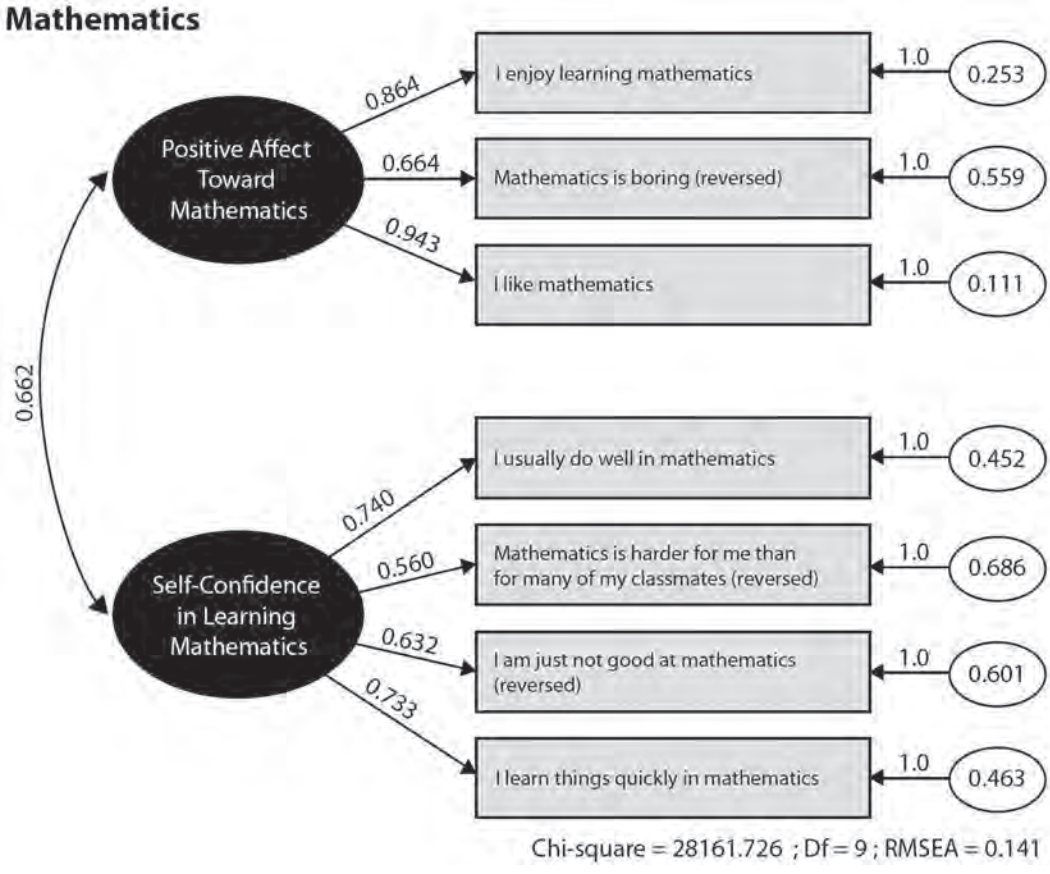


圖4-4-2　TIMSS 2007小四模式適配度檢測之結構方程式圖形

資料來源：TIMSS 2007 Technical Report, p.299

2. Root Mean Square Residual（RMR）

RMR＜0.05表示適配度可接受。

3. Comparative Fit Index（CFI）

0.90＜CFI＜0.95表示適配度可接受；CFI＞0.95表示適配度良好。

4. Non-normed Fit Index（NNFI）

0.90＜NNFI＜0.95表示適配度可接受；NNFI＞0.95表示適配度良好。

表4-4-13為整理PISA 2003針對各種學習策略進行模式適配度檢測與潛在相關之數據分析比較。

（二）TIMSS

使用驗證性因素分析，參考指標如下：

1. Chi-square值

雖然分析有提供卡方值進行模式試配度檢測使用，但是因為受樣本數影響，僅提供當作參考，主要參考指標仍以RMSEA為主。

2. Root-Mean Square Error of Approximation（RMSEA）

RMSEA值小於0.10表示模式適配度為合理的、可接受。圖4-4-2為TIMSS 2007模式適配度檢測之結構方程式示意圖。

（三）綜合討論與建議

表4-4-14對PISA、TIMSS與TASA在模式適配度檢定方法上進行分析比較，可以發現PISA及TIMSS皆選用驗證性因素分析（CFA），驗證性因素分析可用來確認資料模式是否如研究者所預期的形式，藉以檢驗潛在變項存在與否、評估測驗之信效度與檢測特定理論架構下的因素結構。參考指標以Root-Mean Square Error of Approximation（RMSEA）等等為主，軟體選擇上則較彈性，LISREL或Mplus皆可，此可供TASA進行問卷指標分析時參考使用。

表4-4-14　問卷之模式適配度檢定比較

PISA	TIMSS	TASA
1. 驗證性因素分析 2. LISREL軟體 3. 參考指標：RMSEA、RMR、CFI、NNFI。	1. 驗證性因素分析 2. Mplus軟體 3. 參考指標：RMSEA、Chi-square檢定	無此資料

以TASA資料為例，圖4-4-3為使用TASA某年度某縣市有關學習策略之資料進行模式適配度檢測所匯出之結構方程式示意圖形，選用軟體為LISREL，建議TASA後續可針對問卷指標進行更多模式適配度檢定實徵資料的試驗，提高問卷分析之穩定度。

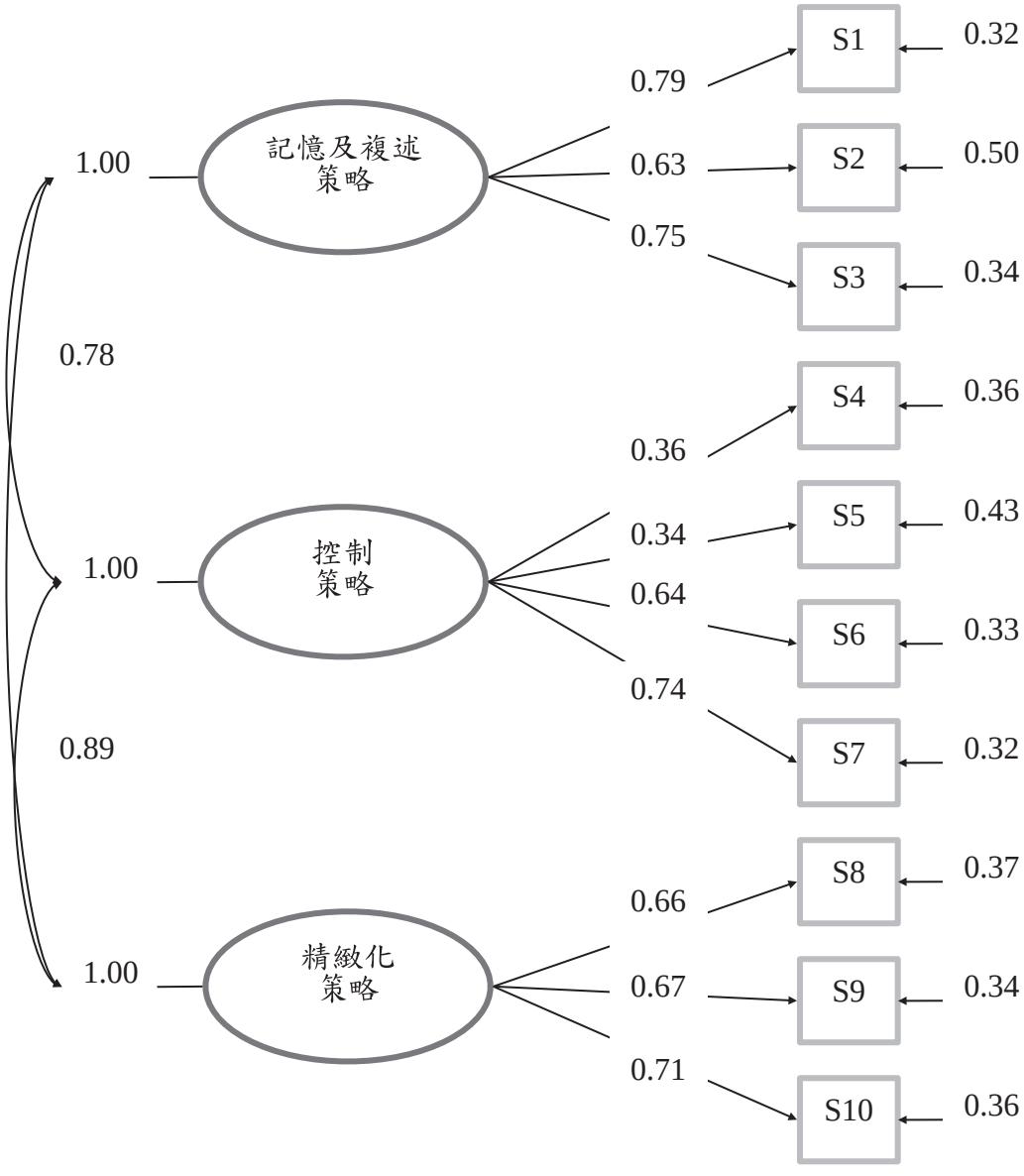


圖4-4-3　TASA有關學習策略指標模式適配度檢測之結構方程式示意圖

表4-4-15為整理2007年TASA學生問卷中適合進行CFA分析之問卷指標，括弧內為該潛在變項指標所包含之題目數量，可提供往後欲進行模式適配度檢測參考使用。

表4-4-15　TASA問卷中適合進行CFA檢測之題項指標整理

社經地位	父親教育程度(1)、母親教育程度(1)、文化資源(10)、文藝課程(10)、文化活動(3)
親子關係	(4)
同儕關係	(4)
師生關係	(4)
班級常規	(4)
學習策略	記憶複述(3)、控制(4)、精緻化(3)
學習偏好	合作(5)、競爭(4)

( )：代表題項數量

第五節　量尺化程序

教育測驗有兩個主要目的，目的一為測量特定學生的知識和技能，學生的表現關乎他或她的未來（職業、入學等），因此降低個體誤差的估計是非常重要的。目的二是評量一個群體的知識和技能，個體的表現將不會影響他們的學校職業或專業生活，這種情況下，降低對目標群體推論的誤差將比降低個體層級的測量誤差更重要，國際的大型測驗是屬於第二個目的（von Davier, Gonzalez, & Mislevy, 2009）。大型測驗的目標是從一群具有代表性的樣本，收集特定內容領域的知識或技巧的相關資料，關注群體的進展情形，因此大型測驗有興趣的統計量數是群體的平均數、標準差、某些層級表現的百分比、百分位數，和關連於上述統計量數的標準誤。

當測驗的內容廣泛，常使用矩陣抽樣設計（matrix-sampling design），使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，由於學生只使用題庫內的一部分試題測量，個體能力的測量會伴隨著相當程度的測量誤差，傳統的個體能力的估計方法，像是最大概似估計法（maximum likelihood estimation, MLE）、期望後驗法（expected a posterior, EAP）這種對於學生個別能力的估計提供最佳的點估計的方式，並不適用於群體能力的估計，集合個體的能力值估計群體的特性將會產生嚴重的偏誤。（Foy, Galia, &Li, 2008; Lee, Grigg & Dion, 2007 ;OECD, 2005）。

目前NAEP、TIMSS和PISA等大型測驗是以試題反應理論（item response theory,IRT）為基礎，透過多重插補，也就是「可能值方法」（plausible values methodology）（Allen,Carlson,Johnson,&Mislevy,1999;Foy, Galia, &Li, 2008; OECD,2005）進行量尺化程序，可能值方法是以潛在迴歸模式，加入學生背景變項計算後驗分佈，並抽取可能值，以利於次級資料分析者使用，可能值方法中沒有先估計個體的能力再計算群體參數，而是使用可得的資料，包含學生的答題反應和背景變項資料直接估計母群的參數，可以使參數的估計較準確（Mislevy& Sheehan, 1989）。以下將以NAEP1998（Allen, Carlson, Johnson, & Mislevy,1999），PISA2003（OECD, 2005）和TIMSS2007（Foy, Galia,& Li, 2008）的技術報告為主，針對這三大測驗與國內目前正在進行之TASA所使用的量尺化程序作一整理說明。

壹、測驗實施與題本設計

不同測驗之量尺化程序的方式會依據測驗實施目的與測驗題本設計的不同而有所差異，因此，要瞭解測驗之量尺化程序，首先必須知道測驗之實施目的與題本設計的方式。

一、測驗實施

NAEP測驗設計共有三種，包含：（1）主要測驗（main assessment）；（2）長期趨勢測驗（assessment for long-term trend）；（3）特殊測驗（special assessment）。PISA主要是測量國際間15歲學生在閱讀、數學與科學之知識能力。每隔三年定期施測此三項領域，並且從中挑出一項當作主要領域，其餘兩項則為次要領域，2000主要領域為閱讀，2003年為數學，2006年為科學。TIMSS主要測量學生之數學與科學能力，從1995年開始每隔四年定期施測，每位學生皆會接受到數學與科學兩項領域。TASA施測對象為臺灣4、6、8年級與高中職二年級學生，施測科目包含國語文、英語文、數學、自然、社會，以評量臺灣學生學習成就表現。以下就三大測驗與TASA之測驗實施年級與時間、測驗科目整理成表4-5-1，表4-5-2作一綜合性比較。

表4-5-1　測驗實施年級與時間之綜合比較

NAEP	主要測驗：4、8、12年級學生 長期趨勢測驗：9、13、17歲學生
PISA	15歲學生，每三年舉辦一次
TIMSS	4、8年級學生，每三年舉辦一次
TASA	4、6、8年級與高中職二年級 2005年施測6年級 2006年施測4、6、8年級與高中職二年級 2007年施測4、6、8年級與高中職二年級 2009年以後每年施測順序依序為4與6年級、8年級、高中職二年級

表4-5-2　測驗科目之綜合比較

NAEP	主要測驗：數學、科學、閱讀、寫作、社會學、美國歷史、公民、地理、文學、音樂、美術、電腦技能等 長期趨勢測驗：閱讀與寫作、數學與科學等
PISA	每次評量會從數學、科學與閱讀三種領域中選出一種為主要領域、其他兩種為次要領域。 2000年主要領域為閱讀 2003年主要領域為數學 2006年主要領域為科學
TIMSS	數學與科學
TASA	2005年施測科目為國語文、英語文、數學 2006年以後施測科目為國語文、英語文、數學、社會、自然 ※社會科小四不施測 ※2009年以後英語文小四不施測

二、題本設計

當題本設計──集中式BIB設計（focused balanced incomplete block）在1988年提出後，BIB與PBIB（partially balanced incomplete block）等化設計之變化類型一直沿用至今。BIB設計與PBIB設計之題本皆由試題組合成之區塊連結組合而成，故學生不需作答太多的試題，僅是接受部份試題區塊，且藉由學生接受部份相同之試題區塊來連結題本。而集中式BIB與PBIB設計提供特定科目之估計，集中式（focused）指的是題本由相同科目之試題區塊組合而成，即每位學生接受相同之受測科目（Nancy, James & John, 2001）。

（一）BIB設計

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（Kuehl, 2000；曾玉琳、王暄博、郭伯臣、許天維，2005），亦即題庫中的每個試題所受測的學生約為相同的。

（二）PBIB設計

PBIB是由Bose & Nair（1939）提出，在此設計中各試題區塊出現次數需相等，但是成對試題區塊的出現次數是不完全相同的，亦即某些成對試題區塊的出現次數是多過於其餘成對試題區塊的出現次數。平衡設計在需要的不完整區塊無法建構出每種實驗情境，使得各區塊需要出現的重複數可能變成過高，題本也隨之增加，但是部分平衡設計在較少的重複數下便可架構出來。

為了清楚呈現大型測驗之題本設計，以下將舉例說明大型測驗之題本設計：

1. NAEP

   以NAEP1998年技術報告提到之4年級公民為例。公民使用之題本設計為BIB設計，而1998年4年級公民使用的是6個試題區塊組合成18個題本之BIB設計變化類型，其變化的原因是為了讓試題區塊在題本前後出現次數一致，故將第16到18個題本設計為將13到15個題本的兩個試題區塊作交換後組成（Andrew & Terry, 2001）。以表4-5-3作說明。

表4-5-3　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
B1	M1	M2	B10	M4	M6
B2	M2	M3	B11	M5	M1
B3	M3	M4	B12	M6	M2
B4	M4	M5	B13	M1	M4
B5	M5	M6	B14	M2	M5
B6	M6	M1	B15	M3	M6
B7	M1	M3	B16	M4	M1
B8	M2	M4	B17	M5	M2
B9	M3	M5	B18	M6	M3

資料來源：NAEP 1998 Technical Report, p.408

2. PISA

PISA2006年使用每個題本包含4個試題區塊，每個試題區塊在題本中出現次數4次，成對試題區塊在各題本中出現次數1次之BIB設計（OECD, 2009），表4-5-4為PISA2006年之題本區塊設計。

表4-5-4　PISA 2006年題本區塊設計表

題本	區塊 I	區塊 II	區塊 III	區塊 IV
B1	S1	S2	S4	S7
B2	S2	S3	M3	R1
B3	S3	S4	M4	M1
B4	S4	M3	S5	M2
B5	S5	S6	S7	S3
B6	S6	R2	R1	S4
B7	S7	R1	M2	M4
B8	M1	M2	S2	S6
B9	M2	S1	S3	R2
B10	M3	M4	S6	S1
B11	M4	S5	R2	S2
B12	R1	M1	S1	S5
B13	R2	S7	M1	M3

註：M代表數學，S代表科學，R代表閱讀

資料來源：PISA 2006 Technical Report, p.29

3. TIMSS

每個題本由四個試題區塊組合而成（每個題本均包含數學與科學各兩個試題區塊），而為了連結不同題本，每個試題區塊在題本中出現2次（Graham, Christine, Alka, & Ebru, 2008）。表4-5-5為TIMSS2007年之題本區塊設計。

表4-5-5　TIMSS 2007年題本區塊設計表

題本	區塊（Part I）		區塊（Part II）	
B1	M01	M02	S01	S02
B2	S02	S03	M02	M03
B3	M03	M04	S03	S04
B4	S04	S05	M04	M05
B5	M05	M06	S05	S06
B6	S06	S07	M06	M07
B7	M07	M08	S07	S08
B8	S08	S09	M08	M09
B9	M09	M10	S09	S10
B10	S10	S11	M10	M11
B11	M11	M12	S11	S12
B12	S12	S13	M12	M13
B13	M13	M14	S13	S14
B14	S14	S01	M14	M01

註：M代表數學，S代表科學

資料來源：TIMSS 2007 Technical Report, p.34

4. TASA

題本設計方面，由於英語文包含聽讀與寫說，故採取每個題本由相同試題區塊組成之題本排列設計，為了連結各題本，試題區塊在不同題本出現不只一次；數學在2005、2006年使用PBIB設計；2007年施測科目除了英語文外，其餘科目皆使用NEAT設計；除此之外，其餘年段與科目皆使用BIB設計。表4-5-6為TASA2009年數學科4年級之題本區塊設計（每個題本由3個試題區塊組合而成，共計13個試題區塊組合成26個題本之BIB設計）。

表4-5-6　TASA 2009年數學科4年級題本區塊設計表

題本序號	區塊I	區塊II	區塊III	題本序號	區塊I	區塊II	區塊III
S1	M11	M10	M1	S14	M4	M1	M12
S2	M6	M11	M8	S15	M6	M10	M13
S3	M2	M12	M6	S16	M13	M3	M7
S4	M10	M9	M7	S17	M8	M9	M12
S5	M7	M11	M2	S18	M4	M2	M10
S6	M7	M6	M4	S19	M3	M5	M6
S7	M8	M7	M1	S20	M5	M8	M10
S8	M1	M6	M9	S21	M2	M5	M9
S9	M12	M13	M11	S22	M9	M13	M4
S10	M12	M7	M5	S23	M5	M1	M13
S11	M9	M3	M11	S24	M13	M8	M2
S12	M10	M12	M3	S25	M11	M4	M5
S13	M1	M2	M3	S26	M3	M4	M8

資料來源：TASA 2009 數學科成果報告（頁12，未出版）

從上述之綜合分析可以發現，不同的大型測驗使用不同的題本設計，整理如表4-5-7。

表4-5-7　題本區塊設計之綜合比較

NAEP	1996年技術報告中指出各科目之題本區塊設計為 數學：BIB設計 科學：BIB設計 閱讀：PBIB設計 寫作：PBIB設計
PISA	BIB設計
TIMSS	每個題本由四個試題區塊組合而成（每個題本均包含數學與科學各兩個試題區塊），而為了連結不同題本，每個試題區塊在題本中出現2次
TASA	除英語文與2005、2006年數學外，2005年、2006年、2009年其餘科目皆為BIB設計 2005年數學為PBIB設計，2006年數學為BIB設計 2007年除英語文外為NEAT設計 英語文因為包含聽讀與寫說，故採取每個題本由相同試題區塊組成之題本設計

NAEP的主要測驗和TASA一個題本僅涵蓋一種領域，NAEP1996年的長期追蹤測驗（NAEP1998沒有實施長期追蹤測驗）、PISA和TIMSS一個題本可能涵蓋不同領域，整理如表4-5-8。

表4-5-8　題本領域之綜合比較

NAEP	主要測驗：一個題本僅一種領域 長期趨勢測驗：一個題本涵蓋閱讀與寫作、數學與科學（1996）
PISA	題本由閱讀、數學與科學三種不同領域組成，題本可能僅只有一個領域或是三種領域皆有涵蓋
TIMSS	題本均涵蓋數學與科學兩種領域
TASA	一個題本僅一種領域

三、綜合討論與建議

測驗實施對象與施測科目而言，TASA測驗實施目的與NAEP較為相似，再加上TASA測驗實施目的為已確定的既有目標，因此本書並無針對TASA測驗實施目的進行建議之處。

而題本設計方面，NAEP使用BIB或PBIB設計，PISA使用之題本設計為BIB設計，TIMSS考量到學生需接受數學與科學兩種領域之測驗，故題本排列採取數學與科學之試題塊分別放置在題本前、後位置順序。就施測目的與BIB設計之優點，也考量到TASA目前使用題本設計之方法，其實與NAEP、PISA同為BIB設計，所以建議TASA在題本設計繼續使用BIB設計。

貳、可能值方法

目前在NAEP1998、PISA2003和TIMSS2007的技術報告中，學生的成就資料是以「可能值」的資料型態提供給次級資料的分析者。在NAEP1998、PISA2003和TIMSS2007的技術報告中說明，試題反應模式中，個體的能力值是觀察不到，即個體能力的測量含有不確定性，這些不確定性在計算群體統計量和相關連的標準誤時，應被考量。可能值是每一位學生的不可觀察特質（能力）的多重差補值，這些多重差補值能反應個體能力估計的不確定性，可能值最早是在1983-1984年NAEP的資料分析中被使用。表4-5-9整理列出各技術報告中所描述可能值的使用時機。

表4-5-9　可能值的使用時機

NAEP	PISA	TIMSS
IRT模式是用來測量個體的能力，當個題被施測的題數夠多時（60題以上），個體能力測量的不確定性可以被忽略。但這種方式在大型測驗中卻不可行，原因是：受試者只被施測相對較少的試題、測驗的形式（題數、題型、試題的內容）不一樣，可能值可用來解決這些困境。	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。	矩陣抽樣設計，學生只使用題庫內的一部分試題測量，個體能力的測量會伴隨著相當程度的測量誤差，透過個體能力估計母群的參數會有偏誤，可能值是解決此一問題之一方法。

一、可能值方法的理論

可能值是在給予學生的答題反應和相關條件變項的條件下，呈現學生可能合理的能力值範圍，不是直接估計每一位學生的能力值，而是估計一位學生能力值的機率分佈，即後驗分佈，從後驗分佈中隨機抽取學生的可能值。在NAEP1998、TIMSS2007和PISA2003的技術報告中，可能值理論推導的公式大致是相同的，只是數學符號定義和條件變數的定義不大一樣，茲以PISA2003的可能值理論推導為例說明。

試題反應模式為條件機率的模式，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$f_\theta(\theta;\alpha)$。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佈的母體，其平均數為μ，變異數為σ^2。也就是：

$$f_\theta(\theta;\alpha) = f_\theta(\theta;\mu,\sigma^2) = (2\pi\sigma^2)^{-1/2}\exp\left[-\frac{(\theta-\mu)^2}{2\sigma^2}\right] \qquad (4.4.1)$$

或者同義的式子，

$$\theta = \mu + E \qquad (4.4.2)$$

其中，$E \sim N(0,\sigma^2)$。

Adams（1997）等人使用迴歸模式$Y_n^T\beta$取代平均數μ，其中Y_n是一個u的向量，對於學生n，Y_n是條件變數，是固定且是已知，β是一個相對應的迴歸係數向量。例如，Y_n可以由性別或社經地位等學生變項所構成。則學生n的母群模式可表示為：

$$\theta_n = Y_n^T\beta + E_n \qquad (4.4.3)$$

其中，假設 $E_n \overset{iid}{\sim} N(0,\sigma^2)$。$E_n$ 的分佈應該會和 θ_n 相同，只是將其轉換為平均數為0，利用迴歸模式 $Y_n^T\beta$ 取代平均數 μ，其中 Y_n 為u的矩陣，β 為迴歸係數。例如 Y_n 可以被視為學生的性別、社經地位或者主修的科目，則母體的模式可以被替換為如下：

$$f_n(\theta_n;Y_n,\beta,\sigma^2)=(2\pi\sigma^2)^{-1/2}\exp[-\frac{1}{2\sigma^2}(\theta_n-Y_n^T\beta)'(\theta_n-Y_n^T\beta)] \quad (4.4.4)$$

為一常態分配，平均數為 $Y_n^T\beta$，及變異數為 σ^2，若使用公式（4.4.4）估算母體分配，則需要估算的參數為 β，σ^2 和 ξ（試題參數），其邊際後驗機率可以被表示為：

$$h_\theta(\theta_n;Y_n,\xi,\beta,\sigma^2\mid x_n)=\frac{f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;Y_n,\beta,\sigma^2)}{f_x(x_n;Y_n,\xi,\beta,\sigma^2)} \quad (4.4.5)$$

如果是多維度變量母群模式，模式如下：

$$f_\theta(\theta_n;\gamma,\Sigma)=(2\pi)^{-d/2}\mid\Sigma\mid^{-1/2}\exp\left[-\frac{1}{2}(\grave{e}_n(\theta_n-\gamma)^{-1}(\theta_n-\gamma)\right] \quad (4.4.6)$$

其中，γ 是一個 $u\times d$ 的迴歸係數矩陣，Σ 是一個 $d\times d$ 的變異數共變數矩陣，每一位受試者的能力值之後驗分佈，如下所示（Adams, Wilson & Wang, 1997）：

$$\begin{aligned}h_\theta(\theta_n;\xi,\gamma,\Sigma\mid x_n)&=\frac{f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;,\gamma,\Sigma)}{f_x(x_n;\xi,\gamma,\Sigma)}\\&=\frac{f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;\gamma,\Sigma)}{\int_\theta f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;\gamma,\Sigma)}\end{aligned} \quad (4.4.7)$$

表4-5-10　可能值的理論公式

NAEP	PISA	TIMSS
預期條件分佈 (predictive conditional ：distribution) $p(\theta_{\underline{r}}\mid x_{\underline{r}},y_{\underline{r}},\Gamma,\Sigma)$ $\propto p(x_{\underline{r}}\mid\theta_{\underline{r}},y_{\underline{r}},\Gamma,\Sigma)$ $\times p(\theta_{\underline{r}}\mid y_{\underline{r}},\Gamma,\Sigma)$ $= p(x_{\underline{r}}\mid\theta_{\underline{r}})\times p(\theta_{\underline{r}}\mid y_{\underline{r}},\Gamma,\Sigma)$（4.4.8）	$h_\theta(\theta_n;Y_n,\xi,\gamma,\Sigma\mid x_n)$ $=\frac{f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;Y_n,\gamma,\Sigma)}{f_x(x_n;Y_n,\xi,\gamma,\Sigma)}$ $=\frac{f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;Y_n,\gamma,\Sigma)}{\int_\theta f_n(x_n;\xi\mid\theta_n)f_\theta(\theta_n;Y_n,\gamma,\Sigma)}$	$P(\theta_j\mid x_j,y_j,\Gamma,\Sigma)$ $\propto P(x_j\mid\theta_j,y_j,\Gamma,\Sigma)$ $P(\theta_j\mid y_j,\Gamma,\Sigma)$ $= P(x_j\mid\theta_j)$ $P(\theta_j\mid y_j,\Gamma,\Sigma)$
$\theta_{\underline{r}}$：學生 r 的能力向量 $x_{\underline{r}}$：學生 r 的作答反應向量 $y_{\underline{r}}$：學生 r 的條件變數向量 Γ,Σ：迴歸模式的參數 $p(\theta_{\underline{r}}\mid y_{\underline{r}},\Gamma,\Sigma)$ 服從常態分佈	θ_n:學生 n 的能力向量 Y_n:學生 n 的條件變數向量 x_n:學生 n 的作答反應向量 ξ：試題參數 γ,Σ：迴歸模式參數	$P(x_j\mid\theta_j)$：試題反應模式 $P(\theta_j\mid y_j,\Gamma,\Sigma)$：在背景變項 y_j、迴歸參數 Γ 和 Σ 的條件下，學生 j 的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。

在PISA 中，使用公式4.4.7產生每一位學生的可能值。表4-5-10整理列出各技術報告中可能值的理論公式。

大型測驗中，使用背景問卷的資料作為母群回歸模式中的條件變數，由於背景問卷資料變數很多，故在大型測驗中常使用主成分分析的方式，縮減問卷變項再納入條件變數，PISA2003中指出，主成分分析的變數需能解釋原始資料95%的變異，而NAEP1998和TIMSS2007則是要能解釋原始資料90%的變異。在PISA2003和TIMSS2007中設定主要條件變數，主要條件變數是指沒有透過主成分分析，一定會被納入迴歸模式中的條件變數，如性別。表4-5-11整理列出各技術報告中條件變數的定義。

表4-5-11　條件變數的設定

NAEP	PISA	TIMSS
1. 從背景變數中，取出約200個主成分，包含背景變項的主要效果和交互作用效果，能解釋約90%的變異。 2. 對於長期追蹤研究，直接使用個數較少的背景變項的主要效果和交互作用，而不使用主成分分析。	1. 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） 直接視為是主要條件變數 。 2. 將學生問卷中的變數虛擬編碼（dummy coded）。 3. 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。	1. 性別（虛擬編碼）、試卷的語言（虛擬編碼）、學生所隸屬的學校班級（criterion-scales）、特定選擇的國家變數（虛擬編碼）是主要條件變數。 2. 類別變數，使用虛擬編碼。 3. 連續變數（出生年、家裡人口）使用criterion scaling，計算interim achievement score，計算的方式請參閱Beaton（1969）。 4. 使用PCA分析虛擬編碼 和criterion-scales 變數，取可解釋90%變異的成分。

二、可能值的抽取步驟

可能值是從學生的能力後驗分佈中隨機抽取，目前在NAEP1998、TIMSS2007和PISA2003的技術報告中每一位學生抽取5個可能值，以PISA為例，說明可能值抽取的步驟。NAEP1998和TIMSS2007的可能值抽取步驟列於表4-5-11。

NAEP1998和TIMSS2007是使用EM演算法則估計後驗分佈，PISA2003是以對於每一個國家的資料集是合適的試題反應模式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，使用蒙地卡羅積分法，對於每一位學生，從能力值的邊際後驗機率（4.4.7）隨機抽取可能值，表4-5-12整理列出各技術報告中可能值之抽取步驟。

表4-5-12　可能值之抽取步驟

NAEP	PISA	TIMSS
使用EM演算法則估計得到 $\hat{\Gamma},\hat{\Sigma}$ 步驟一： 從平均數是 $\hat{\Gamma}$，變異數是 $\hat{\Sigma}$ 的分佈中抽取一個 Γ。 步驟二： 將抽取到的 Γ 和固定的 $\hat{\Sigma}$ 帶入公式（4.4.8），使用EM演算法，可以計算受試者r的預期條件分佈的 $\overline{\theta_r}$ 和 Σ_r。 步驟三： 從近似多變量常態分佈（$\overline{\theta_r}$ 和 Σ_r）的分佈抽取 $\overline{\theta_r}$。 此三個步驟重複五次，產生五個集合的能力差補值。	步驟一： 對於每一個受試者n，從多變量常態分佈， $f_\theta(\theta_n;Y_n,\Upsilon,\Sigma)$ 產生M vector-valued random deviates, $\{\varphi_{mn}\}_{m=1}^{M}$。 步驟二： 使用蒙地卡羅積分法逼近式子（4.4.7）的分母 $\int_\theta f_x(\mathbf{x};\xi\,|\,\theta)f_\theta(\mathbf{x},\gamma,\Sigma)\,d\theta$ $\approx \frac{1}{M}\sum_{m=1}^{M} f_x(\mathbf{x};\xi\,|\,\varphi_{mn}) \equiv \Im$ 同時，計算 $P_{mn} = f_x(x_n;\xi\,|\,\varphi_{mn})$ $f_\theta(\varphi_{mn};W_n,\gamma,\Sigma)$ $\{\varphi_{mn}, P_{mn}/\Im\}_{m=1}$ 的集合可視為式子（4.4.7）的後驗機率函數之近似。 步驟三： 機率值 φ_{nj} 可藉由以下公式求得： $q_{nj} = \frac{P_{mn}}{\sum_{m=1}^{M} P_{mn}}$ 步驟四: 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可能值： $\sum_{s=1}^{i_0-1} q_{sn} < \eta_i < \sum_{s=1}^{i_0} q_{sn}$	步驟一: 從一個近似常態的分配 $P(\Gamma,\Sigma\,|\,x_j,y_j)$，固定 Σ 為 $\hat{\Sigma}$，抽取一個 Γ。 步驟二: 在Γ的條件下，（且固定 $\Sigma=\hat{\Sigma}$）學生j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。

三、計算可能值之軟體

TIMSS2007使用ETS的MGROUP的軟體產生IRT的能力值，輸入的值是學生的作答反應組型、試題參數和條件變數，輸出值是可能值。PISA2003是使用ACER所發展的ConQuest軟體。NAEP1998是使用EST所發展的MGROUP、CGROUP、BGROUPM和GROUP軟體。MGROUP適用一主題內測量p個量尺，CGROUP：是不同主題測量多個量尺，BGROUP只能測量單一量尺，計算較準確，且因為在E步驟使用numeric quadrature，只能應用於一維或二維向度的能力。如果量尺只有單一向度，MGROUP,CGROUP,BGROUP準確度差不多；如果量尺是多向度，CGROUP和BGROUP對於能力間的相關估計較準確，但BGROUP只能用於雙變數。表4-5-13列出三大測驗所使用之可能值分析軟體。

表4-5-13　可能值分析軟體

NAEP	PISA	TIMSS
MGROUP CGROUP BGROUP	ConQuest	MGROUP

四、使用可能值的資料分析

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，而後推論相對應的母群參數T，可惜的是，θ是未知的。大型測驗中，將θ視為遺失資料並且用條件期望值近似$t(\theta, y)$。

$$t^*(x,y) = E\left|t(\underline{\theta},\underline{y}) \mid \underline{x},\underline{y}\right| = \int t(\underline{\theta},\underline{y})p(\underline{\theta} \mid \underline{x},\underline{y})d\underline{\theta} \qquad (4.4.9)$$

其中$(\underline{\theta},\underline{y}) = (\theta_1, y_1, \theta_2, y_2, \theta_3, y_3, \cdots, \theta_J, y_J,)$，$(\theta_j, y_j)$是學生$j$的能力向量和條件變數。給予學生$j$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，抽樣的變異在TIMSS2007和NAEP1998中是藉由jackknife variance estimation procedure 估計而得，在PISA2003是使用Fay's variant of the Balanced Repeated Replication估計。

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，這樣估計群體時會較準確。當模式被正確界定時，可能值

可以提供群體參數的一致性估計，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）。

可能值可以被用來計算公式（4.4.9）進而得到 T，計算方式如下：

步驟一：使用每一位學生的第一組可能值向量計算 T，就像可能值是 θ 的真值，將結果記為 T_1。

步驟二：計算 T_1 的抽樣變異，也就是第一組可能值的抽樣變異 Var_1。

步驟三：重複步驟一步驟二分別計算第二組~第五組的可能值得到 T_u 和 Var_u，u=2,…5。

步驟四：T 的最佳估計值是從不同集合的可能值平均而得。

$$\hat{T} = \frac{\sum_u T_u}{5}$$

$\hat{T}$ 的變異數估計包含兩個成分：

$$\overline{U} = \frac{\sum_u Var_u}{M}$$

$$B_M = \frac{\sum_u (T_u - \hat{T})^2}{M - 1}$$

$\hat{T}$ 的總變異

$$Var(\hat{T}) = \overline{U} + (1 + M^{-1})B_M$$

在 $Var(\hat{T})$ 的第一個變異成分源自於從母群抽樣學生的不確定性（抽樣變異），第二個成分是抽樣學生的θ 無法準確知道，只能透過x和y間接而得，是屬於測量變異。表4-5-14是三大測驗所使用之標準誤計算公式。

表4-5-14　標準誤計算公式

NAEP	PISA	TIMSS
$V = U^* + (1 + M^{-1})B$ *where* $U^* = \sum_{m=1}^{M} \frac{U_m}{M}$ $B = \sum_{m=1}^{M} \frac{(\hat{t}_m - t^*)^2}{(M-1)}$ 抽樣變異：Jackknife estimate	$V = U^* + (1 + M^{-1})B_M$ *where* $U^* = \frac{1}{M}\sum_{m=1}^{M} U_m$ $B_M = \frac{1}{M-1}\sum_{m=1}^{M} (\hat{r}_m - r^*)$ 抽樣變異：Fay's variant of the Balanced Repeated Replication $\sigma^2_{(\hat{\theta})} = \frac{1}{G(1-K)^2} \sum_{i=1}^{G} (\hat{\theta}_{(i)} - \hat{\theta})^2$ $\sigma^2_{(\hat{\theta})} = \frac{1}{80(1-0.5)^2} \sum_{i=1}^{80} (\hat{\theta}_{(i)} - \hat{\theta})^2$ $= \frac{1}{20} \sum_{i=1}^{80} (\hat{\theta}_{(i)} - \hat{\theta})^2$	$Var(\hat{T}) = \overline{U} + (1 + M^{-1})B_M$ $\overline{U} = \frac{\sum_u Var_u}{M}$ $B_M = \frac{\sum_u (T_u - \hat{T})}{M-1}$ 抽樣變異：Jackknife variance estimation procedure

五、建立共同量尺

量尺化程序包含同一施測年度中各題本間的量尺化過程，以及不同施測年度間的量尺化過程，即將該年度之施測資料拉到前一年度的量尺化過程，NAEP、PISA、TIMSS是使用以試題反應理論模式為基礎再加入背景變項之迴歸模式所估計出的「可能值」建立共同量尺，TASA仍是使用以試題反應理論模式為基礎估計而得的「能力值」建立共同量尺。以下將介紹此兩種過程。

（一）同年度題本間量尺化程序

NAEP1998年閱讀與寫作之題本設計使用BIB設計，公民使用PBIB設計，PISA2006年使用BIB設計。BIB與PBIB設計的特點在於試題區塊在不同題本間出現次數相同，故不同題本間可以透過相同之試題區塊來作連結。NAEP1998年技術報告中提到，不同科目間之參數估計採取各別估計，且採用同時估計法將該科所有試題連結到同一量尺上。而PISA在技術報告中找不到同年度題本間量尺化相關文獻。

TIMSS2007年之各題本包含4個試題區塊，為了連結各題本，每個試題區塊在題

本中出現次數為2次。在TIMSS2007年技術報告中提到，題本間之連結數學與科學皆採取同時估計法來將試題連結到同一量尺上。

TASA除英語文外，其餘科目在2005年、2006年、2009年皆使用BIB或PBIB設計，2007年使用NEAT設計。雖然題本設計在不同年度間有所改變，但是在同年度題本間量尺化過程皆採取同時估計法將所有試題拉到同一量尺上。

（二）不同年度間量尺化程序

1. NAEP

NAEP在不同年度間量尺化程序是經由共同校準連結到之前的測驗，藉由共同群體（common population）與線性轉換（linear transformation）來將兩次測驗連結在一起（NAEP Technical Documentation, 2009）。連結步驟如下：

(1)得到前一次測驗之報告測度（reporting metric）。

(2)將兩次測驗資料結合在一起作同時估計以得到一個臨時性量尺（provisional scale）。

(3)由臨時性量尺中挑選出前一次測驗之PVs。

(4)透過線性轉換法將前一次測驗之臨時性量尺轉換至前一次測驗之報告測度上。

(5)使用相同之線性轉換法將該測驗之臨時性量尺轉換成報告測度。

連結之後之量尺分數平均數接近0，再轉換成平均數150、標準差35或是平均數250、標準差50之報告測度。

線性轉換公式如下：

$$Z_{reporting} = Z_{provisional}$$

$$\frac{PV_r - \mu_r}{\sigma_r} = \frac{PV_p - \mu_p}{\sigma_p}$$

$$PV_r - \mu_r = \frac{\sigma_r}{\sigma_p} PV_p - \frac{\sigma_r}{\sigma_p}\mu_p$$

$$PV_r = \frac{\sigma_r}{\sigma_p} PV_p + \mu_r - \frac{\sigma_r}{\sigma_p}\mu_p$$

$$令\ A = \frac{\sigma_r}{\sigma_p},\ B = \mu_r - A\mu_p$$

$$PV_r = A\ PV_p + B$$

2. PISA（TIMSS在不同年度間量尺化方法與PISA相同）

連結PISA 2000年與PISA 2003年閱讀與科學之步驟（OECD, 2009）：

(1)將PISA 2000年OECD會員國施測資料在固定2003年之定錨試題參數後重新估計。

(2)將25個會員國資料結合在一起來計算每個領域之平均數與標準差，此時各國的權重是相同的。

(3)將步驟2之平均數與標準差與PISA 2000年報告測度之平均數與標準差作比較，並且將PISA 2003年之量尺分數透過線性轉換到PISA 2000年。

3. TASA

TASA在不同年度間量尺化程序是採取固定試題參數法，並透過同時估計法將所有試題一起估計得到新的年段之量尺分數。步驟如下：

(1)取得該年度與前一次測驗之定錨試題試題參數。

(2)藉由固定前一次測驗估計而得之定錨試題參數來估計該年度測驗。

(3)由步驟2獲得連結之後之參數值轉換成平均數250、標準差50之量尺分數。

TASA在不同年度間量尺化程序，如果以實證資料2006小四數學、2007小四數學為例來說明，其流程如下圖4-5-1：

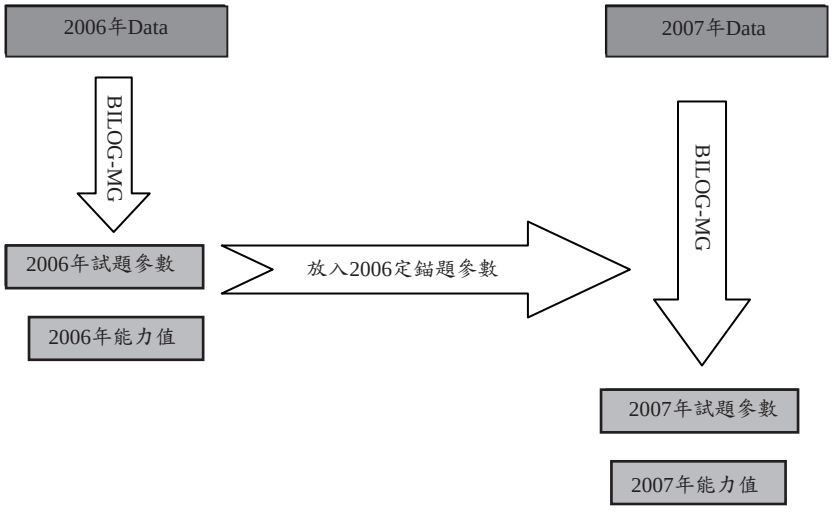


圖4-5-1　TASA不同年度間量尺化過程

各大型測驗所使用的量尺分數範圍不大一樣，整理如表4-5-15。

表4-5-15　量尺分數範圍之綜合比較

NAEP	數學：平均數250，標準差50 科學、閱讀、寫作：平均數150，標準差35
PISA	平均數500，標準差100
TIMSS	平均數500，標準差100
TASA	平均數250，標準差50

六、綜合討論與建議

在建立量尺已經探討了NAEP、PISA、TIMSS與TASA之同年度間量尺化程序及不同年度間量尺化程序，目前TASA的資料分析並未使用可能值方法建立量尺，透過上述之文獻探討，建議可能值分析流程與建立共同量尺的過程。

（一）可能值方法

TASA的施測採BIB等化設計，受試者只需接受若干試題區塊的試題，且不同受試者可能接受部分相同、完全相同、或完全不同的試題區塊。其優點在於能進行大量的施測試題（如可高達195題），可含括較廣的內容領域，但又不會造成受試者之精神負荷（因為每位受試者僅接受約45題），除學科能力之評量外，納入學校、學生背景問卷，能探討如社經地位高低、子女數（如探討少子化影響）、單親家庭、動機、情緒等心理狀態對學業成就之影響等，測驗的資料可對個人或團體進行追蹤比較（臺灣學生學習成就評量資料庫電子報，2009）。由於學生只接受某些試題區塊測量，個體能力的測量會伴隨著相當程度的測量誤差，且TASA也有背景變項之問卷，建議使用可能值方法計算可能值，除了能得到群體統計量的良好估計值外，藉由可能值資料的釋出，能提供給次級資料分析者進行學生學習成效相關因素之探討，促進相關教育議題的討論與連結。

1. 可能值方法的理論

可能值方法是透過潛在迴歸模式，加入學生背景變項計算後驗分佈，理論的模式可參閱PISA2003的技術報告，背景變項的建立方式為參考NAEP（Beaton, 1987）、TIMSS（Macaskill, Adams & Wu, 1998）和PISA（OECD, 2005）。建立步驟如下所示：

步驟一：將性別、社經背景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。

步驟三：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90%的變異。

步驟四：使用已校正的試題參數，和經由主成分分析得到的條件變數估計群體參數分佈。

步驟五：使用上述的方法抽取五個可能值向量。

2. 可能值的抽取步驟

可能值抽取方式如文獻中說明。

3. 產生可能值的軟體

由於ConQuest軟體，具有視窗化界面且有詳細的操作手冊可供參考，建議以ConQuest軟體作為分析軟體。

4. 可能值的資料分析

標準誤的計算須包含測量變異和抽樣變異。透過五個可能值可計算測量變異，在抽樣變異建議以Jackknife variance estimation procedure為主。

5. 建立共同量尺

同年度間量尺化程序方面，由於目前僅探討了同年度題本間量尺化程序以及NAEP在不同年度之量尺化程序、PISA閱讀與科學連結2000年與2003年之量尺化程序。研究發現NAEP與TIMSS在同年度題本間量尺化程序皆採取同時估計法，此方法與目前TASA之方法一致。因此建議TASA在同年度間量尺化方法一樣繼續採用同時估計法來進行量尺化。

在不同年度間量尺化程序方面，由於NAEP是將兩年度資料放在一起進行同時估計，再經由線性轉換進行等化，如此考量到TASA為每年施測，如果採行此方法進行不同年度間量尺化，會造成資料量過大，較不適用於TASA上，而PISA、TIMSS使用分開估計法比較適用於TASA資料上，因此建議TASA使用PISA在不同年度間量尺化方法進行等化，其中以2006、2007年小四數學為例，如下圖4-5-2。

(1) 將TASA2006年與2007年小四數學分開估計，取得各年度學生能力PV值、試題參數。

(2) 將2007年定錨題試題參數放入2006年資料中，將2006年資料重新估計出新的2006年學生能力PV值。

(3) 將2007年學生能力PV值利用線性轉換方法轉換成新的2007年學生能力PV值。

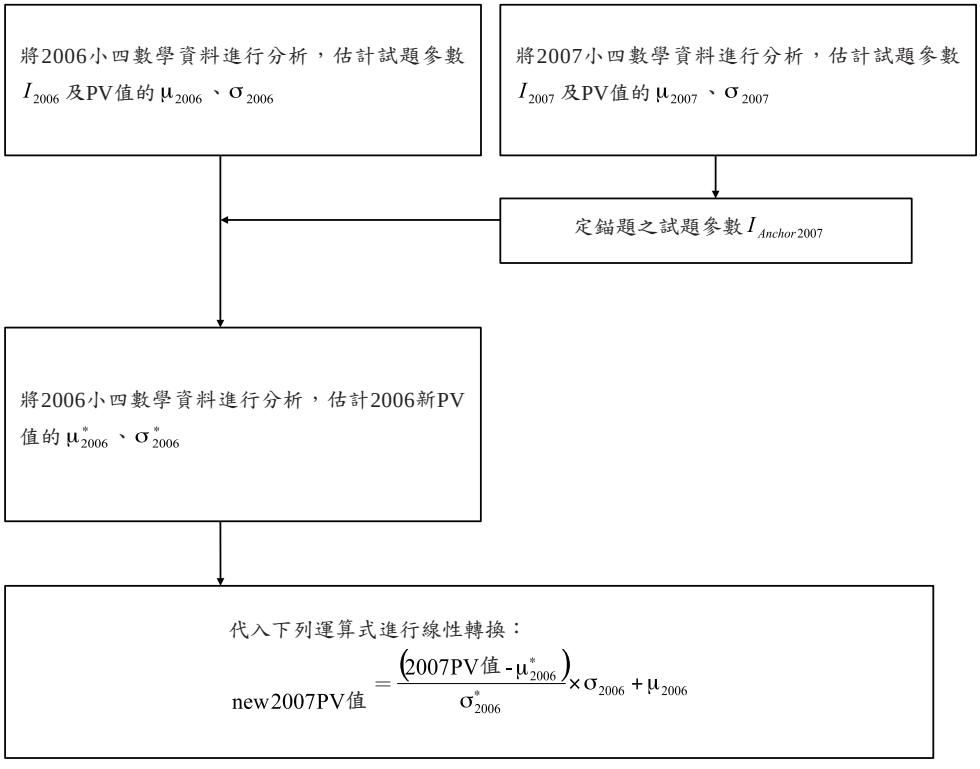


圖4-5-2　TASA實徵資料中不同年度間量尺化之過程

參、建立試題圖（item map）

試題圖主要是顯示各試題在量尺上的分佈，讓讀者可以清楚的看到各題在難度上的排序與差異，因此本書針對NAEP、PISA、TIMSS、TASA探討是否有建置試題圖（item map），並探討試題圖中所含有的訊息。表4-5-16為各大型測驗試題圖之比較。以下針對NAEP、PISA、TASA試題圖內容進行描述如下：

表4-5-16　各大型測驗中試題圖之比較

	NAEP	PISA	TIMSS	TASA
試題圖	1. 結合量尺分數 2. 將量尺分數區分為進階、精熟、基礎 3. 分為MC與CR試題 4. 針對題目進行描述	1. 結合量尺分數 2. 如為部份給分試題則將各難度類別皆列出在試題圖中 3. 針對題目進行描述	無此資料	1. 結合量尺分數 2. 將量尺分數區分為進階、精熟、基礎 3. 皆為MC試題（因計算量尺分數時並無納入CR試題） 4. 針對題目進行描述

一、NAEP

試題圖主要是藉由量尺上各試題的位置來說明每一年級量尺學生在各科目領域的認知與完成程度；而試題在量尺上的位置代表學生可能答對該題。

NAEP 試題圖主要在重點在答對試題所需的知識與技能（knowledge and skills）。以MC試題來說，即為答對試題的機率；CR試題則為試題的分數層級。NAEP所使用之4選項MC試題為74%的答對機率，5選項MC試題為72%的答對機率，CR試題則為65%獲得分數層級的機率。由於CR試題有不同的分數層級，所以CR試題在item map上不只有一個位置（NAEP Technical Documentation, 2009）。

圖4-5-3為NAEP2009四年級數學的試題圖，其中內容上分類為數的特性與運算、測量、幾何、資料分析統計與機率、代數五個內容。在這個試題圖中，有標示出進階層級的量尺分數282、精熟層級的量尺分數249、基礎層級的量尺分數214，並且標示出各個技能相對應試題的類型是屬於MC試題或者是CR試題。而在MC試題中的「在圖表中找到中位數」這個技能需要量尺分數300才有較高的機率可以答對，而這個技能是屬於在進階的量尺分數以上。在每個CR試題中皆會有對於此題的分數層級描述，如「正確」、「部分正確」、「不正確」這些層級上的描述，如下圖4-5-3中，在「格子上畫點以滿足給定的條件」這個技能有四種分數層級，在量尺分數為277時分數層級為具延伸性的（Extended）、量尺分數為263時分數層級為滿意的

（Satisfactory）、量尺分數為260時分數層級為部份的（Partial）、量尺分數為230時分數層級為最小的（Minimal）。

二、PISA

從PISA科學評量中的資料去估計出試題難度，其中某些試題為部份給分試題，而在建立試題圖的過程必須將部份給分試題的各個難度類別視為不同的難度，以PISA科學評量為例，題目共有103題，但難度類別有109個，因為其中有部份給分的試題。（OECD, 2009）

在試題發展的過程中，專家學者試圖對每個試題進行質性分析並且針對每個試題所需具備的認知需求觀點進行描述，每個題目被決定需要具備哪些能力與知識類型才能夠正確回答。其中也定義了各題對於科學的相關是個人的、社會的或全球性的。（OECD, 2009）

圖4-5-4是PISA2006中部份科學試題的試題圖，其中每一列為不同試題，而在這些被選出的部份試題依據了試題難度在試題圖上被排序，其中難度最難的被排序在試題圖的最上面，難度最低的被排序在最下面，而在試題圖中提供了相對應的量尺分數以及試題需求描述。

圖4-5-3　2009年NAEP數學試題圖

資料來源：NAEP Item Map: Mathematics, Grade 4, 2009（http://nces.ed.gov/nationsreportcard/itemmaps/）

Code	Item name	Item difficulty on PISA scale	Item demands	Competency			Knowledge						Focus		
							of				about				
				Identifying scientific issues	*Explaining phenomena scientifically*	*Using scientific evidence*	*Physical systems*	*Living systems*	*Earth and space systems*	*Technology systems*	*Scientific enquiry*	*Scientific explanations*	Personal	Social	Global
S485Q05(2)	ACID RAIN	717	The reason for a control in an investigation is understood and explicitly recognised. An ability to understand the modelling in the investigation is a pre-requisite.	•							•		•		
S114Q05	GREENHOUSE	709	There is a pre-requisite to understand the need to control variables. Knowledge of factors contributing to the greenhouse effect is then applied in determining a variable to be controlled.		•				•						•
S114Q04(2)	GREENHOUSE	659	Given a conclusion can compare two graphs and locate corresponding areas that are at odds with that conclusion and accurately describe that difference.			•						•			•
S447Q05	SUNSCREENS	616	Correctly interprets a dataset expressed diagrammatically and provides an explanation that summarises the data.			•						•	•		
S447Q02	SUNSCREENS	588	The control 'aspects' of an investigation are recognised.	•							•		•		
S493Q05	PHYSICAL EXERCISE	583	Recognition that increased exercise results in increased respiration and thus the need for more oxygen and/or removal of more carbon dioxide.		•			•					•		
S114Q04(1)	GREENHOUSE	568	Recognises differences in two graphs relating to a phenomenon but cannot provide a clear explanation as to why the differences are at odds with a given conclusion.			•						•			•
S213Q01	CLOTHES	567	Can apply knowledge of the features of a scientific investigation to decisions about whether specific issues are scientifically investigatable.	•							•			•	
S493Q01	PHYSICAL EXERCISE	545	Can identify some features of physical exercise that are advantageous to health – cardiovascular system, bodyweight.		•			•					•		
S114Q03	GREENHOUSE	529	Shows an understanding of what two graphs relating to a phenomenon are depicting and can compare them for similarities.			•						•			•
S485Q05(1)	ACID RAIN	513	Recognises that a comparison is being made between two tests but is unable to articulate the purpose of the control.	•							•		•		
S477Q04	MARY MONTAGU	507	Recognises that the immune systems of young and old people are less resistant to viruses than those of the general population.		•			•						•	
S447Q03	SUNSCREENS	499	Can recognise the change and measured variables from a description of an investigation and as a consequence identify the question motivating the investigation.	•							•		•		
S426Q07	GRAND CANYON	485	Can recognise issues in which scientific measurement can be applied to answwering a question.	•							•			•	
S485Q03	ACID RAIN	460	Recognises that the loss of gas in a chemical reaction results in a reduction of mass for the products left behind.			•	•						•		
S426Q03	GRAND CANYON	451	Applies knowledge that water increases in volume as it changes from liquid to solid.		•				•					•	
S477Q03	MARY MONTAGU	431	Recalls knowledge of the role of antibodies in immunity.		•			•						•	
S508Q03	GENETICALLY MODIFIED CROPS	421	Understands that a fair test involves finding out if an outcome is affected by a range of extraneous conditions.	•							•			•	
S213Q02	CLOTHES	399	Can select the correct apparatus to measure an electric current.		•					•			•		
S493Q03	PHYSICAL EXERCISE	386	Rejects the notion that fats are formed in the muscles and knows that the rate of flow of blood increases during exercise.		•			•					•		

圖4-5-4　PISA 2006中部份科學試題的試題圖

資料來源：PISA 2006 Technical Report, p.291

三、綜合討論與建議

在NAEP、PISA、TASA試題圖中，皆以量尺分數為基準，針對測驗試題進行難度上的排序並將學生相對應的量尺分數對應出來在試題圖上，並且在相對應的題目詳細敘述該題目的知識內容，在NAEP與TASA中更在試題圖上將量尺分數分為進階、精熟、基礎，讓讀者可以清楚的了解試題是屬於在何種層級，NAEP與TASA試題圖中將試題清楚註明是MC試題或CR試題。在PISA的試題圖中，針對各題進行質性的探究，歸納出各題所對應的知識層面。建議TASA可以參考2006年試題圖並納入部分給分模式的CR試題到試題圖中，並且可以加入PISA在製作試題圖時，針對每個試題進行題目本身的質性探究，說明試題所對應的知識及技能。

第六節　出版報告

出版報告為大型測驗成果之展現，其報告類別繁多，多數以成果報告與技術報告為主。首先，本節先就NAEP、PISA、TIMSS、TASA之出版報告類別作一整理，提出版報告類別之建議；其次，在成果報告的部份，選擇形式偏向一般大眾閱讀之「大眾版」NAEP 2007、PISA 2006成果報告，以及報告形式偏向學術研究報告之「學術版」TIMSS 2007、TASA 2007成果報告，針對測驗成果報告的特色與內容，分別作一整理說明（Lee, Grigg, & Dion, 2007；Mullis, Martin,& Foy, 2008；OECD,2007；TASA,2007），提出TASA成果報告之建議；最後，比較NAEP、PISA、TIMSS、TASA技術報告的內容，提出TASA技術報告之建議。

壹、NAEP、PISA、TIMSS、TASA出版報告之類別

一、NAEP

以2007年之NAEP為例，其釋出之出版報告有NAEP 2007數學成績報告（The Nation's Report Card: Mathematics 2007）、NAEP 2007各州數學成績報告（NAEP Mathematics 2007 State Snapshot Reports）、NAEP 2007各城市區域數學成績報告（NAEP Mathematics 2007 District Snapshot Reports），其中並無發現技術報告。出版報告內容如下所述：

（一）NAEP 2007數學成績報告：內容包括測驗架構與設計、施測對象、量尺分數、成果報告。但並沒有如技術報告詳細分析方法論。

（二）NAEP 2007各州數學成績報告：將各州的成果報告分開撰寫成一頁的簡易報告。

（三）NAEP 2007各城市區域數學成績報告：將各城市的成果報告分開撰寫成一頁的簡易報告。

二、PISA

以2006年之PISA為例，其釋出之出版報告有15歲的男生與女生學校表現（Equally prepared for life? How 15-year-old boys and girls perform in school）、PISA 2006在科學上高成就的學生（Top of the Class - High Performers in Science in PISA 2006）、PISA 2006十五歲的學生在環境科學與地球科學的表現（Green at Fifteen? How 15-year-olds perform in environmental science and geoscience in PISA 2006）、PISA測驗中的樣本試題（PISA Take the Test: Sample Questions from the OECD's PISA Assessments）、PISA在SAS與SPSS的資料分析手冊（PISA Data Analysis Manual: SAS and SPSS, Second Edition）、PISA 2006技術報告（PISA 2006 Technical

Report）、PISA 2006科學能力應用在明日的世界（PISA 2006 Science Competencies for Tomorrow's World）、PISA 2006架構下的科學、閱讀、數學素養（Assessing Scientific, Reading and Mathematical Literacy: A Framework for PISA 2006）。出版報告內容如下所述：

（一）15歲的男生與女生學校表現：學校學生仍然有因為性別的因素產生差距，雖然在大部分國家，男生和女生在PISA表現出相同的測驗結果。

（二）PISA 2006在科學上高成就的學生：什麼樣的因素導致學生在科學上的高成就？文中提到學生個人因素、學校因素、教育體制因素如何關係到學生的科學成就。

（三）PISA 2006十五歲的學生在環境科學與地球科學的表現：在OECD的國家中學生對於環境議題的覺察。

（四）PISA測驗中的樣本試題：從PISA中公佈可用的測驗試題，而其中部分試題是從2000、2003、2006資料庫中釋出，而其他的試題繼續發展用在測驗上。

（五）PISA在SAS與SPSS的資料分析手冊：提供了研究者了解PISA資料庫和使用SAS、SPSS來完成分析。

（六）PISA 2006技術報告：詳細提供研究者學習有關PISA2006各項分析的方法論，讓研究者可以有能力操作相同的分析。

（七）PISA 2006科學能力應用在明日的世界：內容展示出最近的PISA資料庫的結果，而焦點放在科學的部份，也評估數學和閱讀的結果。

（八）PISA 2006架構下的科學、閱讀、數學素養：展示出概念架構下的PISA2006。它包含了測驗學生科學素養、數學能力、閱讀能力三種發展中的架構，在每個領域中，都被定義了學生必須具備的內涵、技能與知識，其中也提供了一些實例。

三、TIMSS

以2007年之TIMSS為例，其釋出之出版報告有TIMSS 2007測驗架構（TIMSS 2007 Assessment Frameworks）、TIMSS 2007國際科學報告（TIMSS 2007 International Science Reports）、TIMSS 2007國際數學報告（TIMSS 2007 International Mathematics Reports）、TIMSS 2007大全（TIMSS 2007 Encyclopedia）、TIMSS 2007技術報告（TIMSS 2007 Technical Report）、TIMSS 2007國際資料庫與使用手冊（TIMSS 2007 International Database and User Guide）、TIMSS釋出題目（TIMSS 2007 Released Items）。出版報告內容如下所述：

（一）TIMSS 2007測驗架構：TIMSS 2007 測驗架構包含了數學測驗、科學測驗、問卷背景變相三個架構，也概述了測驗設計。

（二）TIMSS 2007國際科學報告：總結了TIMSS2007在59個參與國家和8個基準國在科學上的成就，也描述了科學教育的背景。它也包含了TIMSS 1995、TIMSS 1999、TIMSS 2003、TIMSS 2007的科學成就趨勢研究。成就測驗結果也顯示科學內涵，和認知領域。報告包含了大量的學生背景及對學生對科學的態度、科學課程、教師的教育與訓練、班級氣氛。

（三）TIMSS 2007國際數學報告：總結了TIMSS2007在59個參與國家和8個基準國在數學上的成就，也描述了數學教育的背景。它也包含了TIMSS 1995、TIMSS 1999、TIMSS 2003、TIMSS 2007的數學成就趨勢研究。成就測驗結果也顯示數學內涵，和認知領域。報告包含了大量的學生背景及對學生對數學的態度、數學課程、教師的教育與訓練、班級氣氛。

（四）TIMSS 2007大全：TIMSS 2007大全描述了參加國家在數學和科學的教學和學習的背景，包括了學校數學和科學課程、教師教學需求、測驗的形式及測驗的使用。

（五）TIMSS 2007技術報告：TIMSS 2007技術報告敘述了整體的設計與施行，內容包含了抽樣、資料蒐集、量尺化、資料分析及報告。

（六）TIMSS 2007國際資料庫與使用手冊：為了促使二手分析來改善四年級和八年級的數學和科學教育，TIMSS 2007釋出能用的資料庫供研究與分析，而這個資料庫是由59個國家和8個基準國的學生、老師、學校的背景資料所組成的。使用手冊描述了TIMSS 2007國際資料庫的內容與格式。

（七）TIMSS 2007釋出題目：釋出四年級與八年級數學和科學之施測試題。

四、TASA

以2007年與2009年之TASA為例，其釋出之出版報告有TASA 2009技術報告、TASA 2007評量結果報告、TASA 2007資料釋出光碟與TASA電子報。出版報告內容如下所述：

（一）TASA 2009技術報告：內容包含測驗設計及發展、問卷的發展、抽樣架構、施測過程、資料庫的建置、項目分析、量尺化程序、學習成就表現描述。

（二）TASA 2007評量結果報告：內容包含抽樣設計、命題架構、正式測驗題本試題分配、評量結果分析、結論與建議。

（三）TASA 2007資料釋出光碟：內容包含問卷變項、抽樣學生作答反應與問卷作答，供研究者進行二手分析。

（四）TASA電子報：於2009年11月1日創刊，形式為短篇報告，以TASA資料庫為分析依據，提供特定議題的統計資訊，以供教育界人士參考與使用。刊物內容規劃含專題、技術報告、專論及新聞專輯。

表4-6-1　NAEP、PISA、TIMSS、TASA出版報告之類別

	NAEP 2007	PISA 2006	TIMSS 2007	TASA2007,2009
技術報告	無	PISA 2006技術報告	TIMSS 2007技術報告	TASA 2009技術報告
背景變項描述	無	無	TIMSS 2007大全	無
成果報告	1. NAEP 2007數學成就報告 2. NAEP 2007各州數學成就報告 3. NAEP 2007各城市區域數學成就報告	1. 15歲的男生與女生學校表現 2. PISA 2006在科學上高成就的學生 3. PISA 2006十五歲的學生在環境科學與地球科學的表現 4. PISA 2006科學能力應用在明日的世界	1. TIMSS 2007國際數學報告 2. TIMSS 2007國際科學報告	TASA 2007評量結果報告
測驗架構	無	PISA 2006架構下的科學、閱讀、數學素養	TIMSS 2007測驗架構	無
資料分析手冊	無	PISA在SAS與SPSS的資料分析手冊	無	無
資料釋出	無	無	TIMSS 2007國際資料庫與使用手冊	TASA 2007資料釋出光碟、TASA電子報
題目釋出		PISA測驗中的樣本試題	TIMSS 2007釋出題目	

五、綜合討論與建議

綜合上述，TASA目前出版報告類別有：技術報告、成果報告、資料釋出光碟以及TASA電子報等四種，尚無以非測統專業人員為閱讀者之成果報告，因此對於出版報告之類別，提出「增加大眾版成果報告」之建議，以期透過大眾版成果報告之呈現，提供教育部長官與國內民眾一個簡要瞭解TASA重點成果之報告。

貳、NAEP、PISA、TIMSS、TASA成果報告之分析與建議

成果報告主要有以民眾為閱讀者之「大眾版」成果報告，以及偏向學術性質的「學術版」成果報告兩種。以下茲就形式偏向一般大眾閱讀之「大眾版」成果報告：NAEP 2007、PISA 2006，以及報告形式偏向學術研究報告之「學術版」成果報告：TIMSS 2007、TASA 2007，針對測驗成果報告的特色與內容，分別作一整理說明；最後，提出綜合討論與建議。

一、大眾版成果報告：NAEP 2007、PISA 2006

NAEP是美國評量學生成就的代表，由國家評量管理委員會（National Assessment Governing Board，以下簡稱NAGB）負責監控與政策推動，NAGB的成員包含聯邦官員、州政府官員、地方官員，以及教育學者專家、家長，特別的是還包含商業領導者與社會一般民眾的成員代表（Lee, Grigg, & Dion, 2007）。因此，NAEP廣為民眾所知的成果報告「The Nation's Report Card」，目的即是使老師、家長和一般社會大眾，了解關於美國評量的重要資訊。其報告內容以大量淺顯易懂的圖表取代學術性的文字，呈現主題明確，例如以「2007男性較女性高2分」為標題，使閱讀者能於短時間內了解美國學生於學科、族群、州際、性別等議題上歷年來的改變。

PISA是由經濟合作暨發展組織（Organization for Economic Co-operation and Development，以下簡稱OECD）所主辦的一項國際性學生評量與比較，主要是針對即將完成義務教育走入社會的15歲在學學生，以情境化、生活化的試題，評量這些學生在閱讀素養、數學素養、科學素養的表現，以便了解他們是否對進入社會成為公民做好準備（Mullis, Martin,& Foy, 2008），其成果報告對各國教育政策的制定與調整，提供甚多具參考價值的數據與資料。PISA報告內容多以學術性的文字敘述，整篇以文字為主，圖表為輔，段落章節則以小方塊區隔，圖表呈現方式較NAEP複雜，讀者需仔細閱讀才能從報告中獲取資料。PISA報告中獨立「品質與公平正義」章節，可見其對此議題之重視。表4-6-2列出NAEP、PISA成果報告之特色。

表4-6-2　NAEP、PISA成果報告之特色

NAEP	PISA
1. 共64頁（含附錄23頁）。 2. 圖文並重，以大量淺顯易懂的圖表取代學術性的文字。 3. 呈現主題明確，使讀者能於短時間內讀到重點。	1. 共54頁（無附錄）。 2. 以文字為主，圖表為輔，段落多以方塊區隔，若不細讀，短時間內無法抓到重點。 3. 圖表呈現較NAEP複雜、需仔細閱讀。 4. 獨有「品質與公平正義」章節。

以下茲就NAEP 2007與PISA 2006成果報告的內容呈現方式，作一整理說明（Lee, Grigg, & Dion, 2007；Mullis, Martin,& Foy, 2008）。

（一）摘要

摘要針對整篇報告做重點式的摘錄，使讀者能迅速掌握該篇報告的重要成果與發現。NAEP之執行摘要（圖4-6-1）以一目了然的圖表呈現「族群比較、州際比較、學生數學能力質性敘述」三項主題，輔以文字敘述，標題文字即呈現下方內容，例如以「15州與特區在四年級與八年級學習成就皆進步」概括州際學生之成就表現；PISA之關鍵報告（圖4-6-2）全篇以文字敘述呈現，段落前方則以小方塊做為區隔，標題文字較不能使讀者迅速掌握下方內容，例如以「科學表現」為標題，讀者可知下方內容為科學表現，但不知表現為何。

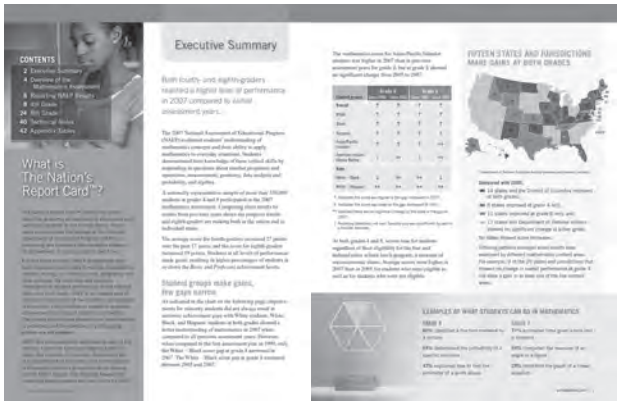


圖4-6-1　The Nation's Report Card（NAEP執行摘要）
資料來源：Lee, Grigg, & Dion (2007：2-3)

圖4-6-2　PISA 2006 Executive Summary（執行摘要）-Key findings
資料來源：Mullis, Martin,& Foy, (2008：3-4)

表4-6-3　NAEP、PISA成果報告之摘要內容

NAEP	PISA
執行摘要 2007年四年級與八年級學生成就表現水平較往年提升 1. 各族群的學生團體進步，少數族群成就斷層變窄。 2. 15州與特區在四年級與八年級學習成就皆進步。 3. 學生數學能力質性敘述。	關鍵報告 1. 科學表現。 2. 閱讀表現。 3. 數學表現。 4. 學生對科學的態度。 5. 學校與層級系統因素。

（二）緒論

NAEP、PISA之緒論內容皆為該測驗之簡介，PISA特別提出該年度測驗的要點與改變，因其為國際測驗，因此亦有參與國家之介紹。

表4-6-4　NAEP、PISA成果報告之緒論內容

NAEP	PISA
什麼是The Nation's Report Card？	PISA 2006 1.背景。 2. PISA 2006要點 3. PISA 2006的改變。 4. 參與的國家。

（三）評量架構與評量設計

NAEP與PISA皆有介紹評量架構與評量設計之章節，NAEP評量架構內涵包括：數字概念與運算、測量、幾何概念、分析與機率、代數，試題則分為低階複雜、中階複雜、高階複雜三等級。PISA此章節亦以表格列出科學素養的測驗內涵，並以實際試題舉例說明七層素養水準之答題情形。

表4-6-5　NAEP、PISA成果報告之評量架構與評量設計內容

NAEP	PISA
數學評量概要 1. 數學評量架構。 2. 評量設計。	科學評量 1. PISA 2006如何測量學生的科學表現？ 2. 科學試題、學生成績與素養水準。 3. PISA科學試題範例。 4. 形成科學議題。 5. 解釋科學現象。 6. 科學舉證。

（四）成就報告

NAEP與PISA之成就報告皆包含與往年之比較、性別比較；因參與對象之不同，NAEP有族群與州際成績之比較，PISA則有各國成績之比較；至於社經背景的影響，NAEP於此章節中包含高社經與低社經學生成績之比較，PISA則另外獨立「品質與公平正義」章節探討。

表4-6-6　NAEP、PISA成果報告之成就報告內容

NAEP	PISA
四年級 1. 成績較前幾年表現都高。 2. 1990-2007不同族群成績。 3. 2007男性較女性高2分。 4. 高社經與低社經學生成績。 5. 各州成績。 6. 評估四年級內容(基本、精熟、進階) 八年級 1. 八年級數學知識增加。 2. 1990-2007不同族群成績。 3. 1990-2007不同性別成績。 4. 高社經與低社經學生成績。 5. 各州成績。 6. 評估八年級內容(基本、精熟、進階)	科學表現 1. 學生科學表現概況。 2. 學生科學素養。 3. 學生的平均表現。 4. 不同國家表現強弱的學生在哪個科學面向有差異？ 閱讀表現 1. PISA 2006閱讀表現與自PISA 2000年來的改變。 2. 閱讀素養。 3. 平均閱讀成績。 4. 自PISA 2000年來的改變。 5. 性別差異。 數學表現 1. PISA 2006數學表現與自PISA 2000年來的改變。 2. 數學素養。 3. 平均數學成績。 4. 自PISA 2003年來的改變。 5. 性別差異。

（五）背景變項

NAEP之背景變項包含樣本、量尺分數、成就水平、試題圖等內容，PISA則包含入學政策、能力分班、鄰近學校選擇權、學校自治、學校資源等等學校問卷之內容。

表4-6-7　NAEP、PISA成果報告之背景變項內容

NAEP	PISA
報告NAEP結果 1. 樣本：2007參與人數。 2. 量尺分數 3. 成就水平：基本、精熟、進階 4. 試題圖。 5. 特殊學生的調適或排除。 6. 解釋結果。	學校與層級系統因素 1. 許可、選取與編組。（入學政策、能力分班） 2. 父母的壓力與鄰近學校選擇權。 3. 附有責任的政策。（給家長的成績是否有與同校相比、與不同學校的相比、國際性標準等三種） 4. 學校自治。（教師僱用權.教師薪資等） 5. 學校資源。（師生比、電腦數等） 6. 教育的品質、公平正義與學生表現之間是有關聯的。

（六）附錄

NAEP成果報告包含技術記錄、附錄表格、其他資訊等附錄，PISA則無附加之附錄。

表4-6-8　NAEP、PISA成果報告之附錄內容

NAEP	PISA
技術記錄（Technical Note） 附錄表格（Appendix Tables） 其他資訊（More Information）	無

（七）其他特有章節

NAEP無特有章節，PISA則有「對科學的態度」、「品質與公平正義」兩章節。

表4-6-9　NAEP、PISA成果報告之特有章節內容

NAEP	PISA
無	對科學的態度 1. 學生對科學參與程度的概況 2. 對科學態度的重要性 3. 評量態度的新方法 4. 學生對科學探究支持嗎？ 5. 學生對科學學習有信心嗎？ 6. 學生對科學有興趣嗎？ 7. 學生會自覺對環境與資源有責任嗎？ 8. 對科學態度的性別差異 品質與公平正義 1. 學校扮演何種不同的角色？ 2. 社經背景的公平正義與學校的品質可能被調解嗎？ 3. 學校與社經背景的型態差異是否指涉不同國家的政策？

二、學術版成果報告：TIMSS 2007、TASA 2007

TIMSS由國際教育成就評鑑協會（The International Association for the Evaluation of Educational Achievement，以下簡稱IEA）所主持，IEA為國際性教育研究機構，主要任務在於針對教育政策與實務進行跨國的比較研究，提供世界各國作為制定教育決策的參考。因此，TIMSS成果報告撰寫方式為正式的學術研究報告，頁數共計473頁，其中包含附錄103頁，多以問句作為標題，目錄即穿插圖表的標題於章節之中，圖表與文字敘述並重，並大量使用正式表格呈現各國比較，第一章至第三章報告數學成就，第四章至第八章報告背景變項。

TASA為我國自2004年起所建置之全國性學生學習成就長期資料庫，其成果報告分為國、英、數、社、自五科，撰寫方式為國內的正式報告格式，與TIMSS之不同處為獨有「結論與建議」章節。

表4-6-10　TIMSS和TASA成果報告之特色

TIMSS	TASA
1. 共473頁（含附錄103頁）。 2. 問句式標題。 3. 圖表與文字敘述並重，包含許多正式的國際比較表，目錄即穿插圖表的標題於章節之中。 4. 第一章至第三章報告數學成就，第四章至第八章報告背景變項。.	1. 共211頁（無附錄）。 2. 撰寫方式為國內的正式報告格式。 3. 獨有「結論與建議」章節。

以下茲就TIMSS 2007與TASA 2007成果報告的內容呈現方式，作一整理說明（OECD,2007；TASA,2007）。

（一）摘要

TIMSS之執行摘要全篇以文字敘述呈現，主題中之段落則於前方加上小三角做為區隔；TASA則無摘要敘述。

Executive Summary

TIMSS 2007 is the fourth in a continuing cycle of international mathematics and science assessments conducted every four years. TIMSS assesses achievement in countries around the world and collects a rich array of information about the educational contexts for learning mathematics and science, with TIMSS 2007 involving more than 60 participants. This report contains the mathematics results for 37 countries and 7 benchmarking participants at the fourth grade and for 50 countries and 7 benchmarking participants at the eighth grade. Trend data are provided at the fourth and eighth grades for those countries that also participated in 1995, 1999, and 2003 (please see the Introduction for more information about TIMSS 2007).

Mathematics Achievement

- At the fourth grade, Hong Kong SAR and Singapore were the top performing countries. They were followed by Chinese Taipei, that had higher average mathematics achievement than all countries except Hong Kong SAR and Singapore, and, in turn, by Japan, that had higher achievement than all of the remaining countries. Kazakhstan, the Russian Federation, England, Latvia, and the Netherlands also performed very well. Several benchmarking participants also had high average mathematics achievement, including the U.S. state of Massachusetts, which performed similarly to Chinese Taipei and the state of Minnesota, which performed similarly to Kazakhstan, the Russian Federation, and England.

- At the eighth grade, Chinese Taipei, Korea, and Singapore had the highest average mathematics achievement. These three countries were followed by Hong Kong SAR and Japan, also performing similarly and having higher achievement than all the other countries except the top three performers. There was a substantial gap in average mathematics achievement between the five Asian countries and the next group of four similarly performing countries, including Hungary, England, the Russian Federation, and the United States. Among the benchmarking participants, the two U.S. states, Massachusetts and Minnesota, and the province of Quebec were outperformed by the five Asian countries but had higher average achievement than the group of four countries. The provinces of Ontario and British Columbia had average achievement similar to the group of four countries.
- Remarkable percentages of students in Asian countries reached the Advanced International Benchmark for mathematics, representing fluency on items involving the most complex topics and reasoning skills. In particular, at the fourth grade, Singapore and Hong Kong SAR had 41 and 40 percent of their students, respectively, achieving at or above the Advanced International Benchmark. At the eighth grade, Chinese Taipei, Korea, and Singapore had 40 to 45 percent of their students achieving at or above the Advanced International Benchmark. The median percentage of students reaching this Benchmark was 5 percent at the fourth grade and 2 percent at the eighth grade.
- Looking at trends across all of the participating countries, not taking into account whether countries have participated in two, three, or four cycles (eighth grade) of TIMSS, more countries showed improvement in average achievement between their first cycle of participation and TIMSS 2007 than declines at the fourth grade, although this was not the pattern at the eighth grade. At the fourth grade, 16 countries had higher average achievement in 2007 than in their first TIMSS assessment, 5 had lower average achievement, and 8 showed no significant change. At the eighth grade, 10 countries had higher average achievement in 2007 than in their initial assessment, 15 lower average achievement, and 13 showed no significant change.

IEA TIMSS & PIRLS

圖4-6-3　TIMSS Executive Summary（執行摘要）
資料來源：OECD (2007：5-6)

表4-6-11　TIMSS和TASA成果報告之摘要內容

TIMSS	TASA
執行摘要 1. 數學成就。 2. 較高數學成就的相關因素。 3. 數學課程與教學。	無

（二）緒論

TIMSS、TASA之緒論內容皆為該測驗之簡介，TASA則於此章節中特別說明抽樣的設計。

表4-6-12　TIMSS和TASA成果報告之緒論內容

TIMSS	TASA
緒論 1. 什麼是TIMSS？ 2. 哪些國家參與TIMSS 2007？ 3. 什麼是TIMSS 2007數學測驗的本質？ 4. 學習數學的背景資訊如何蒐集？ 5. 誰實施TIMSS？	壹、緒論 一、研究緣起。 二、研究目的。 三、研究團隊。 四、抽樣 （一）前言。 （二）分層設計。 （三）實際抽樣。

（三）評量架構與評量設計

TIMSS與TASA皆無獨立介紹評量架構與評量設計之章節，其評量架構與評量設計皆包含在成就報告章節中。

（四）成就報告

因參與對象之不同，NAEP成就報告有國家比較、各國年度改變、各國年段分析，TASA則有不同地理區域、縣市之學生比較；TIMSS有歷年成就之比較，TASA數學科、社會科有歷年成就比較，國語科、英語科、自然科則無。

表4-6-13　TIMSS和TASA成果報告之成就報告內容

TIMSS	TASA
第一章：國際學生數學成就	貳、國小四年級評量結果分析
1. 國家比較。	一、命題架構。
2. 各國年度改變。	二、正式測驗題本試題分配。
3. 各國年段分析。	三、評量結果分析
4. 各國性別差異。	（一）國小四年級學生在國語文科表現水準之界定。
第二章：TIMSS 2007數學成就國際評價標準之表現	（二）國小四年級常模樣本整體表現情形之現況描述。
1. 如何以TIMSS 2007數學成就國際評價標準作國家比較？	（三）國小四年級學生在國語文科不同成就水準的分佈。
2. 四年級-進階評價標準	（四）不同地理區域、縣市之學生在國語文學習成就的差異分析。
3. 四年級-高階評價標準	（五）不同學生背景變項在國語文學習成就表現之差異分析。
4. 四年級-中階評價標準	（六）不同學校背景變項對學生在國語文學習成就表現之差異分析。
5. 四年級-低階評價標準	參、國小六年級評量結果分析
6. 八年級-進階評價標準	（與小四相同）
7. 八年級-高階評價標準	肆、國中二年級評量結果分析
8. 八年級-中階評價標準	（與小四相同）
9. 八年級-低階評價標準	伍、高中二年級評量結果分析
第三章：數學的內容與認知結構成就百分比	（與小四相同）
1. 各國於數學內容與認知結構成就的百分比。	陸、高職二年級評量結果分析
2. 各國於數學內容與認知結構成就的相對強弱。	（與小四相同）
3. 性別差異。	

（五）背景變項

TIMSS之背景變項共分「學生的社經背景與態度對數學的影響」、「數學課程」、「數學教師」、「教室特色與教學」、「學校脈絡的數學學習與教學」五章，TASA之背景變項則含在成就報告章節中。

表4-6-14　TIMSS和TASA成果報告之背景變項內容

TIMSS	TASA
第四章：學生的社經背景與態度對數學的影響 第五章：數學課程 第六章：數學教師 第七章：教室特色與教學 第八章：學校脈絡的數學學習與教學	含在成就報告章節中。

（六）附錄

TIMSS附錄包含「支持文件」、「數學內容與認知結構的平均成績多重比較」、「數學-考試課程的一致性分析」、「數學成就的百分等級與標準差」、「蒙古-數學成就」等國際比較表，多達103頁，TASA則無附錄。

表4-6-15　TIMSS和TASA成果報告之附錄內容

TIMSS	TASA
附錄A：支持文件 附錄B：數學內容與認知結構的平均成績多重比較 附錄C：數學──考試課程的一致性分析 附錄D：數學成就的百分等級與標準差 附錄E：蒙古──數學成就 附錄F：TIMSS 2007組織與個人的責任	無

（七）其他特有章節

TIMSS無特有章節，TASA與其他大型測驗成果報告不同的是獨有「結論與建議」章節，內容包含研究發現與討論、以及給行政機關、學校、教師、家長的建議。

表4-6-16　TIMSS和TASA成果報告之特有章節內容

TIMSS	TASA
無	捌、結論與建議 一、研究發現與討論 二、結論與建議 （一）對行政機關的建議 （二）對學校的建議 （三）對教師的建議 （四）對家長的建議

三、綜合討論與建議

TASA尚無大眾版之成果報告，參考NAEP與PISA之報告，建議TASA大眾版成果報告朝「呈現主題明確，使讀者能於短時間內讀到重點」、「圖文並重，以淺顯易懂的圖表取代學術性的文字」兩特色做規劃。並建議TASA大眾版成果報告之內容包括關鍵報告（當年成就表現重點歸納）、測驗簡介（包括當年度測驗內容的改變）、評量架構與評量設計、成就報告、背景變項、技術附註、附錄表格、其他資訊（機關與作者資訊）。

TASA現有之學術版成果報告，將「命題架構」、「正式測驗題本試題分配」、「表現水準之界定」、「常模樣本整體表現情形之現況描述」、「不同成就水準的分佈」、「不同地理區域、縣市之學生在學科學習成就的差異分析」、「不同學生背景變項在學科學習成就表現之差異分析」與「不同學校背景變項對學生在學科學習成就表現之差異分析」，接放入「各年段評量結果分析」一章中探討。為考量成果探討之嚴整性，建議學術版成果報告將「成就報告」、「評價標準」、「學科內容與認知結構成就百分比」與「背景變項」分開章節做探討，並於緒論前增加「執行摘要」一章，對學術版成果報告做一重點歸納。

表4-6-17　TASA成果報告之建議

大眾版成果報告	學術版成果報告
1. 關鍵報告 （當年成就表現重點歸納） 2. 測驗簡介 （包括當年度測驗內容的改變） 3. 評量架構與評量設計 4. 成就報告 5. 背景變項 6. 技術附註 7. 附錄表格 8. 其他資訊 （機關與作者資訊）	1. 執行摘要 2. 緒論 研究緣起、研究目的、研究團隊、分層設計、實際抽樣 3. 成就報告 (1) 縣市比較 (2) 年度改變 (3) 年段分析 4. 評價標準 5. 學科內容與認知結構成就百分比 6. 背景變項（學生問卷與學校問卷分析） 7. 結論與建議 8. 附錄

參、NAEP、PISA、TIMSS、TASA技術報告之分析與建議

NAEP、PISA、TIMSS、TASA技術報告之內容大致相同，以下茲就四大教育測驗之技術報告內容作一彙整，最後提出綜合討論與建議。

一、NAEP、PISA、TIMSS、TASA之技術報告內容

NAEP、PISA、TIMSS、TASA之技術報告內容，彙整如表4-6-18，其內容大致上可綜合歸納為以下幾個重點：

1. 緒論或概要
2. 測驗設計及發展
3. 問卷的發展
4. 抽樣設計
5. 施測過程
6. 品質管理
7. 檢視權重和抽樣變異
8. 項目分析
9. 量尺化程序
10. 學習成就表現描述
11. 資料庫的建置過程與管理
12. 附錄與參考文獻

表4-6-18　NAEP、PISA、TIMSS、TASA之技術報告內容

內容	NAEP 1998	PISA 2006	TIMSS 2007	TASA 2009
第1章	NAEP 1998的設計與施行概要	PISA 概要	TIMSS 2007大綱	緒論
第2章	發展NAEP 1998閱讀、寫作、公民評量的目標、試題和背景問題	測驗的設計與發展	發展TIMSS 2007數學及科學評量和給分簡介	測驗設計及發展
第3章	全國性評量的抽樣設計	PISA 背景問卷的發展	發展TIMSS 2007背景問卷	問卷的發展
第4章	州評量的抽樣設計	抽樣設計	翻譯成各國適用的TIMSS 2007評量與問卷	抽樣架構
第5章	實地操作與資料收集	測驗的翻譯和文化適切以及檢閱的工具	TIMSS 2007抽樣設計	施測過程
第6章	處理評量工具	實地操作	TIMSS 2007操作過程	資料庫的建置
第7章	專業的計分	品質保證	TIMSS2007 資料收集的品質管理	項目分析

表4-6-18　NAEP、PISA、TIMSS、TASA之技術報告內容（續）

內容	NAEP 1998	PISA 2006	TIMSS 2007	TASA 2009
第8章	資料庫建立、資料輸入的品質控制與資料庫成果	檢視權重和抽樣變異	建立和檢查TIMSS 2007資料庫	量尺化程序
第9章	NAEP1998資料分析	量尺化PISA的認知資料	TIMSS 2007抽樣權重和參加比率	學習成就表現描述
第10章	全國性評量權重過程與抽樣權重變異	資料管理的過程	檢視TIMSS 2007試題統計	-
第11章	州評量的權重過程與變異估計	抽樣結果	TIMSS2007數學和科學評量資料的量尺化	-
第12章	量尺化過程	量尺化的結果	建立TIMSS 2007 背景指標	-
第13章	假設性考驗和NAEP結果的報告	編碼與註記信度研究	TIMSS 2007 數學和科學學生成就的國際標準	-
第14章	1998年全國性與州的閱讀評量架構和工具	資料宣告	-	-
第15章	全國性與州閱讀評量的資料分析介紹	精熟量尺的結構	-	-
第16章	全國性閱讀評量的資料分析	量尺化的過程與背景問卷資料的建構效度	-	-
第17章	州閱讀評量的資料分析	態度量尺的建構效度	-	-
第18章	1998年全國性與州的寫作評量架構和工具	國際資料庫	-	-
第19章	全國性與州寫作評量的資料分析介紹	-	-	-
第20章	全國性寫作評量資料分析	-	-	-
第21章	州寫作評量的資料分析	-	-	-
第22章	1998年公民評量架構和工具	-	-	-
第23章	公民評量的資料分析介紹	-	-	-
第24章	公民評量的資料分析	-	-	-
	附錄	附錄	附錄	-
	參考文獻	-	-	參考文獻

二、綜合討論與建議

綜合以上NAEP、PISA、TIMSS、TASA之技術報告內容分析發現，NAEP 1998、PISA 2006與TIMSS 2007之技術報告中，對於「品質管理」、「檢視權重和抽樣變異」皆有獨立章節探討，因此建議TASA技術報告增加「品質管理」、「檢視權重和抽樣變異」兩章節以及「附錄」。

表4-6-19　TASA技術報告之建議

技術報告
1. 緒論
2. 測驗設計及發展
3. 問卷的發展及建構效度
4. 抽樣設計
5. 施測過程
6. 品質管理
7. 檢視權重和抽樣變異
8. 項目分析
9. 量尺化程序
10. 學習成就表現描述
11. 資料庫的建置過程與管理
12. 附錄
13. 參考文獻

第五章　結論與未來研究方向

第一節　結論

一、抽樣設計與抽樣權重

本書藉由國外大型測驗抽樣設計之探討，可知大型測驗主要皆是透過多階段的抽樣方法抽取受試樣本，主要分成兩個階段，包含受試學校與受試學生的選取。因此，檢視現行TASA的抽樣設計是否有需改善的部分，經研究發現TASA抽樣設計推估抽樣權重時，若不考慮城鄉層級的計算，最終權重總合與母群體受試人數相比有低估的情況發生（以TASA 2006年數學科為例）。由於TASA樣本學校被抽取的機率並沒有依據PPS的抽樣方式，才造成學生最終權重有高估或低估的情況發生，因此，本書建議應調整樣本學校的權重（即調整樣本學校被抽取的機率），以正確估計受試者權重。

是故，本書建議TASA使用二階段分層抽樣設計（國中與國小的部分），第一階為分層叢集隨機抽樣，根據縣市與班級數兩個變項進行分層；第二階段再根據所抽取的樣本學校，以學生個人為抽樣單位進行簡單隨機抽樣。且因應2010年台灣行政地區之改變，以及為了進行縣市或地區性受試樣本成就表現之比較，將依據五都、四個地理區、離島地區劃分PSU。透過此抽樣設計，並採用PISA的Fay複製方法估計抽樣變異。

二、測量模式

（一）選擇題部分，建議TASA繼續使用三參數對數模式，題組題的部分，使用題組模式進行分析；開放性試題部分，目前TASA針對開放性試題僅進行敘述統計分析，未來建議使用一般化部分給分模式進行試題分析。

（二）模式適合度評估方法，TASA目前尚未使用任何方法進行模式適合度評估，建議參考NAEP、TIMSS使用圖形化判斷方式進行模式適合度評估。

三、試題資料

選擇題部分，NAEP、TIMSS及TASA皆採用三參數對數模式，因此TASA目前之分析方法可持續沿用，但是對於開放性試題僅提供敘述統計資料，數據略嫌不足，需進一步進行更深入的試題參數估計，以便獲得更多試題訊息。另外，針對開放性試題進行二次評分之評分者間一致性檢定，為目前TASA所缺乏的，因此TASA於之後之正式施測中，對每一開放性試題應要求至少經過兩位閱卷者評分，如此有助於

進行二次評分之評分者間一致性檢測，提高TASA測驗評量之信度。DIF分析部份，TASA問卷背景調查中包含性別、國籍原生性及慣用語言等資料，因此未來有必要挑選部分變項進行全面性之差異試題功能分析探討，以降低試題在族群間產生作答反應差異之情況發生。

四、問卷背景變項資料

TASA現行分析以三參數對數模式為主，學生能力值為EAP期望後驗估計所得，加上在抽樣上並無抽樣權重之設計，因此分析上並無考慮抽樣誤差及測量誤差，此與NAEP、PISA及TIMSS有極大之差異。因此，增加抽樣權重與可能值之計算，藉以降低抽樣誤差及測量誤差，為TASA目前需改進之方向。另外，問卷題項分析前之信、效度分析乃是基本原則，對各國際大型資料庫而言，皆為不可或缺之一環，因此，TASA針對問卷背景變項分析，可考慮增加Cronbach's alpha信度分析與CFA建構效度分析。

五、量尺化程序

（一）測驗實施與題本設計

在題本設計，考量到TASA目前使用題本設計之方法，其實與NAEP、PISA同為BIB設計，所以建議TASA在題本設計繼續使用BIB設計。

（二）可能值方法

可能值方法是透過潛在迴歸模式，加入學生背景變項計算後驗分佈，從後驗分佈中隨機抽取可能值，建議TASA應以可能值方法進行量尺化程序，步驟如下所示：(1)將性別、社經背景直接視為是條件變數。(2)將學生問卷中的變數虛擬編碼。(3)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90%的變異。(4)使用已校正的試題參數，和經由主成分分析得到的條件變數估計群體參數分佈。(5)使用上述的方法抽取五個可能值向量。

建議TASA在同年度間量尺化方法一樣繼續採用同時估計法來進行量尺化，而TASA在不同年度間量尺化過程上，建議採用PISA在不同年度間量尺化方法進行等化。

（三）建立試題圖

TASA2006年小四數學有針對樣本試題進行試題圖的繪製，但在2009年並無相關的試題圖內容，因此建議TASA可以在成果報告中加入試題圖的分析，使得TASA更加完善，並在試題圖中加入MC與CR試題的區別，讓部份給分的試題也能夠加入分析中，而且可以針對每個試題進行題目本身的質性探究，說明試題所對應的知識及技能。

六、出版報告

（一）增加大眾版成果報告

TASA目前尚無以非測統專業人員為閱讀者之成果報告，因此提出增加大眾版成果報告之建議，並朝「呈現主題明確，使讀者能於短時間內讀到重點」、「圖文並重，以淺顯易懂的圖表取代學術性的文字」兩特色做規劃，以期透過大眾版成果報告之呈現，提供教育部長官與台灣民眾一個簡要瞭解TASA重點成果之報告。

TASA大眾版成果報告之內容包括關鍵報告（當年成就表現重點歸納）、測驗簡介（包括當年度測驗內容的改變）、評量架構與評量設計、成就報告、背景變項、技術附註、附錄表格、其他資訊（機關與作者資訊）。

（二）學術版成果報告內容調整

TASA現有之學術版成果報告，將「命題架構」、「正式測驗題本試題分配」、「表現水準之界定」、「常模樣本整體表現情形之現況描述」、「不同成就水準的分佈」、「不同地理區域、縣市之學生在學科學習成就的差異分析」、「不同學生背景變項在學科學習成就表現之差異分析」與「不同學校背景變項對學生在學科學習成就表現之差異分析」，接放入「各年段評量結果分析」一章中探討。

為考量成果探討之嚴整性，建議學術版成果報告將「成就報告」、「評價標準」、「學科內容與認知結構成就百分比」與「背景變項」分開章節做探討，並於緒論前增加「執行摘要」一章，對學術版成果報告做一重點歸納。

（三）技術報告內容調整

綜合NAEP、PISA、TIMSS、TASA之技術報告內容分析發現，NAEP 1998、PISA 2006與TIMSS 2007之技術報告中，對於「品質管理」、「檢視權重和抽樣變異」皆有獨立章節探討，因此建議TASA技術報告增加「品質管理」、「檢視權重和抽樣變異」兩章節以及「附錄」。

（四）資料分析手冊

PISA提供了SAS與SPSS的資料分析手冊，使研究者了解PISA資料庫和使用SAS、SPSS來完成分析，建議未來TASA提供資料分析手冊給次集資料分析者，提供正確的資料分析程序與方法，避免次級資料分析者誤用。

第二節　未來研究方向

一、TASA自2005年始進行施測，即未探討目前抽樣設計將受試樣本權重值視為相同與忽略抽樣變異的估計等可能產生等問題，針對本書建議之TASA抽樣設計權重值的計算，以及原先TASA抽樣設計權重值的修正等問題，將是後續應該繼續探討的部分。

二、模式適合度之評估方面，大多數的研究均是以單一試題檢驗模式適合度，未來可探討針對整份試卷之模式適合評估。

三、以TASA的評量架構而言，應是屬於多向度能力之評量架構，未來可探討多向度試題反應理論模式應用於TASA資料分析之可行性。

四、題本設計上，可將TASA2006年BIB設計、TASA2007年NEAT設計與TASA2009年BIB設計量尺化後，探討目前TASA現有資料上題本設計在可能值量尺化方法上之影響。

五、可能值方法主要以單向度試題反應理論或多向度單參數試題反應理論為基礎，考量TASA之評量架構與測驗題型，未來可探究以多向度三參數試題反應理論為基礎之可能值方法。

六、可能值方法主要為納入背景變項的考慮，使得回復群體參數更為準確，未來期望進行模擬研究，探討有無納入背景變項，對回復群體參數的影響，並佐以TASA實徵資料進行探討。

七、目前TASA的統計考驗並未納入標準誤的計算，未來建議計算TASA的標準誤，並利用其標準誤進行統計考驗分析。

八、針對開放性試題，TASA未來研究可探討NAEP、TIMSS所使用二參數對數模式、一般化部份給分模式與PISA所使用的多向度隨機係數多項洛基模式進行參數估計與試題特性分析上之差異與適用性，做為自身參考採用依據。

九、問卷背景變項分析部份，TASA未來研究方向可增加探討在單向度IRT與多向度IRT理論架構下，DIF分析、Cronbach's alpha信度分析與CFA建構效度分析是否存在差異性。

十、在建立共同量尺上，考量到PISA與TIMSS使用相同的等化方法，建議TASA也採用相同等化方法，未來研究可進行TASA實證資料不同年度間量尺分數的估算，並探討運用可能值進行等化後之差異。

十一、在TASA2006年並非每科都有針對試題進行試題圖的分析，所以未來研究可以針對TASA2006、2007年、2009年、2010年的資料進行試題圖的分析，讓TASA資料更加完善。

十二、TASA目前出版報告類別有：技術報告、成果報告、資料釋出光碟以及TASA電子報等四種，加上本書所建議增加的大眾版成果報告，TASA出版報告共有：技術報告、成果報告（大眾版）、成果報告（學術版）、資料釋出光碟以及

TASA電子報等五種，以國內需求而言尚屬完善。未來之研究方向，建議整理國際上使用NAEP、PISA、TIMSS資料完成之研究報告，以及其研究題目與成果，以供國內學術研究發展之參考。

參考文獻

中文部分

吳明隆（2006）。**SPSS統計應用學習實務（3版）**。臺北市：知城。

曾玉琳、王暄博、郭伯臣、許天維（2006）。不同BIB設計對測驗等化的影響。*測驗統計年刊*，**第十三輯下期**，209-229。臺中市：國立臺中教育大學。

張郁雯（2008）。對比效應對學業自我概念之影響—發展的觀點。**教育心理學報，40（1）**，23~38。

譚克平（2009）。TIMSS國際教育評比研究簡介。**大型教育資料庫及相關議題學術研討會**。國立臺中教育大學。

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。

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。

臺灣學生學習成就評量資料庫電子報（2009）。檢索日期：2010年1月20日。網址：http://tasa.naer.edu.tw/uploadfiles/file/TASAePaper

英文部分

Adams, R. J., Wilson, M., & Wang, W.-C. (1997). The multidimensional random coefficients multinomial logit model. *Applied Psychological Measurement. 21*, 1-23.

Allen, N. L., Donoghue, J. R., & Schoeps, T. L. (2001). *The NAEP 1998 technical report.* Washington, DC: National Center for Educational Statistics.

Allen N. L.,Carlson J. E., Johnson E. G. ,& Mislevy, R. J.(1999) *The NAEP 1998 technical report.* Educational Testing Service.

Andrew R. W. & Terry L. S., (2001). *The NAEP 1998 Technical Report (NCES 2001-509).* National Assessment Governing Board, U.S. Department of Education.

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.

Bose R. C. & Nair K. R. (1939). *Partially balanced incomplete block designs, Sankhya, 4*, 337-372.

Cohen, J. (1968). Weighted kappa: Nominal scale agreement with provision for scaled disagreement or partial credit. *Psychological Bulletin, 70(4)*, 213-220.

Dorans, N., & Kulick, E. (1986). Demonstrating the utility of the standardization approach to assessing unexpected differential item performance on the Scholastic Aptitude Test. *Journal of Educational Measurement, 23,* 355-368.

Foy p., Galia J.,& Li L.(2008).Scaling the data from the TIMSS 2007 Mathematics and Science assessments.In John F. Olson,Michael O. Martin ,Ina V.S. Mullis.(Eds). *TIMSS 2007 Technical Report.* TIMSS & PIRLS International Study Center,Lynch School of Education, Boston College.

Graham J. R., Christine Y. O’ S., Alka A., & Ebru E. (2008). *TIMSS 2007 Technical Report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Huynh, H. (1994). *Some technical aspects of standard setting.* Paper presented at the Joint Conference on Standard Setting for Large-Scale Assessments. Washington, D.C.

Huynh, H. (1998). On score locations of binary and partial credit items and their applications to item mapping and criterion-referenced interpretation. *Journal of Educational and Behavioral Statistics, 23(1),* 38-58.

Kuehl, R. O. (2000). Design of Experiments: *Statistical Principles of Research Design and Analysis.* CA: Duxbury

Press.

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. (1980). *Applications of Item Response Theory to Practical Testing Problems.* New Jersey: Lawrence Erlbaum.

Martin, M. O. and Kelly, D. L. (1996) *Third International Mathematics and Science Study Technical Report, Volume I: Design and Development.* Chestnut Hill, MA: Boston College.

Mislevy, R. J. & Bock R. D. (1982)*Implementation of the EM algorithm in the estimation of item parameters: The BILOG computer program.* Item Response Theory and Computerized Adaptive Testing Conference , Wayzata, MN, July 27-30, 1982.

Mislevy, R. J. & Sheehan, K. M. (1980). Information matrices in latent-variable models. *Journal of Educational Statistics 14(4),* 335-350.Mislevy, R. J. & Sheehan, K. M. (1987). Marginal estimation procedures, in A.E. Beaton (ed.). The NAEP 1983-1984 Technical Report (Report No. 15-TR-20). Educational Testing Service, Princeton, N.J.

Mislevy, R. J., & Sheehan, K. M. (1989). Information matrices in latent-variable models. *Journal of Educational Statistics, 14,* 335-350.

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.

Martin, M. O., Mullis, I.V.S., & Chrostowski, S. J. (Eds.) (2004). *TIMSS 2003 Technical Report.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Mantel N., & Haenszel, W. M. (1959). Statistical aspects of the analysis of data from retrospective studies of disease. *Journal of the National Cancer Institute, 22,* 719-748.

Mullis, I. V. S., Martin, M. O., Gonzalez, E. J., & Chrostowski, S. J. (2004). *TIMSS 2003 International Mathematics Report: Findings from IEA's Trends in International Mathematics and Science Study at the Fourth and Eighth Grades.* Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

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.

Mantel, N. (1963). Chi-square tests with one degree of freedom: Extensions of the Mantel-Haenszel procedure. *Journal of the American Statistical Association, 58,* 690–700.

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/

Nancy L. A., James E. C., & John R. D. (2001). *The NAEP 1998 Technical Report (NCES 2001-509).* National Assessment Governing Board, U.S. Department of Education.

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

OECD (2006). *Assessing Scientific, Reading and Mathematical Literacy.* OCED, Paris. http://www.oecd.org/document/33/0,3343,en_32252351_32236191_37462369_1_1_1_1,00.html

OECD (2006). *PISA RELEASED ITEMS - MATHEMATICS.* Retrieved February 27, 2009, from http://www.pisa.oecd.org/dataoecd/14/10/38709418.pdf

OECD (2007). *PISA 2006 Science Competencies for Tomorrow's World.* OCED, Paris. Retrieved January 11,

2010, from http://www.pisa.oecd.org/document/2/0,3343,en_32252351_32236191_39718850_1_1_1_1,00.html

OECD (2007). *The programme for international student assessment –PISA.* OECD, Paris. Form http://www.oecd.org/dataoecd/15/13/39725224.pdf

OECD (2009). *Top of the Class - High Performers in Science in PISA 2006.* OCED, Paris. http://www.oecd.org/document/51/0,3343,en_32252351_32236191_42642227_1_1_1_1,00.html

OECD (2009). *Green at Fifteen? How 15-year-olds perform in environmental science and geoscience in PISA 2006.* OCED, Paris. http://www.oecd.org/document/22/0,3343,en_32252351_32236191_42466966_1_1_1_1,00.html

OECD (2009). *PISA Data Analysis Manual: SPSS and SAS, Second Edition.* OCED, Paris. http://www.oecd.org/document/38/0,3343,en_32252351_32236191_42609254_1_1_1_1,00.html

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

OECD (2009). *Equally prepared for life? How 15-year-old boys and girls perform in school.* OCED, Paris. http://www.oecd.org/document/51/0,3343,en_32252351_32236191_42837811_1_1_1_1,00.html

Rock, D. A. (1991). *Subscale dimensionality.* Paper presented at the meeting of the Design and Analysis Committee of the National Assessment of Educational Progress, Washington, DC.Rubin,D. B. (1987).Multiple imputation for nonereponse in surveys.New York:Wiley.

Shealy, R., & Stout, W. (1993). A model-based standardization approach that separates true bias/DIF from group ability differences and detects test bias/DIF as well as item bias/DIF. *Psychometrika, 58,* 159-194.

Wang, W., & Wilson, M. (2005). Exploring local item dependence using a facet random-effects facet model. *Applied Psychological Measurement, 29,* 296-318.

Wang, X., Bradlow, E. T., & Wainer , H. (2004). *User's guide for SCORIGHT (version 3.0): A computer program for scoring tests built of testlets including a module for covariate analysis.* Princeton, NJ: Educational Testing Service; Philadelphia, PA: National Board of Medical Examiners.

Von Davier Matthias, Gonzalez Eugenio,& Mislevy R. J.(2009).What are plausible values and why are they useful ? *IERA Monograph Series:Issues and Methodologies in Large-Scale Assessment,2,* pp.9-36.

Wu, M. L.,R. J. Adams and M. R. wilson(1997), *ConQuest:Multi-Aspect Test Software* [computer program],Australian Council for Education Research,Camberwell.

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

Yates, F. (1936). A new method of arranging variety trials involving a large number of varieties. *J. Agric. Sci. 26,* 424-455.

Zimowski, M. F., Muraki, E., Mislevy, R. J., & Bock, R. D. (2003). *BILOG-MG: Multiple-group IRT analysis and test maintence for binary for binary items.* Mooresvilk IL: Scientic Software.

國家圖書館出版品預行編目（CIP）資料

大型標準化測驗建置流程應用於TASA之研究 / 王暄博等作 ; 郭伯臣, 曾建銘, 吳慧珉主編. -- 初版.
-- 新北市 : 國家教育研究院, 民100.12
面 ; 公分
ISBN 978-986-03-1384-0（平裝）

1.教育測驗 2.資料庫 3.標準化

521.3029 100028211

書　名：大型標準化測驗建置流程應用於TASA之研究
發行人：吳清山
主　編：郭伯臣、曾建銘、吳慧珉
作　者：王暄博、曾筱倩、周慧玲、杜雨潔、蔡翰征、張宛婷、蘇少祖
出版機構：國家教育研究院
地　址：新北市三峽區三樹路2號
網　址：http://www.naer.edu.tw
電　話：（02）8671-1111
出版年月：中國民國101年1月
版　次：初版
其他類型版本說明：無
定　價：新臺幣 200 元
美　編：朱墨形象設計廣告有限公司
印　刷：上校基業有限公司
展　售：政府出版品展售中心
五南文化廣場：臺中市中山路6號
電話：04-22260330；傳真：04-22258234
網址：http://www.wunan.com.tw/
國家書店松江門市：臺北市松江路209號1樓
電話：02-25180207；傳真：02-25180778
網址：http://www.govbooks.com.tw/

GPN：1010100071
ISBN：978-986-03-1384-0（平裝）