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

第三節 TIMSS 大型測驗之探討

TIMSS 主要目的為進行學生數學與科學教育成就趨勢調查研究，測試對象為 4 年級與 8 年級之學生，欲評估學生能否掌握參與社會所需的知識與技能，並藉由國際評比來比較參與地區或國家的教育成效。自 1999 年進行 TIMSS-R 評量後，IEA 計畫每隔四年辦理國際數學與科學教育成就研究一次，並改名為 TIMSS。以下將簡要說明 TIMSS 實施時幾個重要之技術層面（Martin, Mullis, & Chrostowski, 2004）。

壹、評量架構、測驗設計與問卷之發展

一、評量架構

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

二、測驗設計

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

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

三、背景問卷

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

貳、抽樣設計與抽樣權重

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

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

參、試題分析

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

肆、量尺化程序

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

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

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

1.可能值方法簡介

y：所有抽樣學生背景資料的反應

θ: 預估計的能力

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

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

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

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

Γ：背景變數的回歸係數矩陣

Σ：殘差共變異矩陣

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

$P(x_j \mid \theta_j))$：試題反應模式

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

2.條件(Conditioning)

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

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

ε 是常態分布，平均數是 0，變異數是 Σ

Γ 是一矩陣每一欄是每一個能力量尺的效果(effects)

Σ 是量尺之間的殘差變異矩陣。

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

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

3.產生能力值(generating proficiency scores)

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

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

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

這三個步驟重複五次，每一位學生產生 5 個 θ_j 的差補值。

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

4.條件變數

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

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

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

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

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

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

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