TIMSS國際教育評比研究簡介

譚克平／國立台灣師範大學科學教育研究所教授

國際數學與科學教育成就趨勢調查研究（TIMSS）在國際比較教育這個領域中佔有舉足輕重的地位，該研究所開發或採用的多項技術與措施 ，其中包括：抽樣設計、翻譯程序、題本設計、資料品管、資料分析、報告的形式等等，其研究方法有一定程度的嚴密性，因此漸漸成為其它大、小型調查研究參考的對象。從某一個角度來說，該研究以及其前身可說是開啟了近代國際教育評比活動的先端。

一、國際教育評比研究的簡史

二十世紀五十年代後期，一群有志於比較教育的研究者齊聚於德國漢堡，共同商討如何有效進行學校及學生的評鑑，其中一個目的是希望能透過跨國比較的方式，瞭解什麼因素會與學生的學習成就有關，從而探討如何可以改善教學的效果。1967年，由多位對國際比較教育的學者所發起的國際教育成就調查委員會（International Association for the Evaluation of Educational Achievement，簡稱IEA）正式成立，並先後於1964年及1970至1971年分別舉辦了第一次國際數學研究（First International Mathematics Study，簡稱FIMS）及第一次國際科學研究（First International Science Study，簡稱FISS）。此外，在1980至1982年及1984年又舉辦了第二次國際數學研究（Second International Mathematics Study，簡稱SIMS）及第二次國際科學研究（Second International Science Study，簡稱，SISS）。自此國際間大型教育相關的評比研究開始蓬勃發展，例如在1991年美國教育測驗服務社（Educational Testing Service）曾舉辦國際數學及科學教育評鑑計畫（International Assessment of Educational Progress，簡稱IAEP），當年臺灣有參與該研究計畫。及至1994至1995年間，IEA為了延續前兩次的數學與科學的評比研究，特別以綜合的形式舉辦了第三次國際數學與科學教育成就研究（Third International Mathematics and Science Study，簡稱TIMSS），並且以數學及科學的課程為評比的主要依據。是次參與研究的國家或地區為數不少，為了不希望要等每十年左右才進行一次國際評比，再加上希望能快一點追蹤學生學業成就的改變，IEA於四年後（即1999年）再進行了一次評比研究，並稱之為TIMSS-Repeat（簡稱TIMSS-R），而臺灣亦正式參與了該次的研究計畫。爾後IEA希望能將該評比研究正規化成為每四年舉行一次，遂將該研究的名稱更改為國際數學與科學教育成就趨勢調查研究（Trends in International Mathematics and Science Study），簡稱仍維持為TIMSS，而學術界習慣上會對從1995年以降的評比研究稱之為TIMSS 1995、TIMSS 1999、TIMSS 2003等等，以資區隔。另一方面，由於對評比的若干理念不相同，經濟合作暨發展組織（Organisation for Economic Co-operation and Development，簡稱OECD）自2000年開始另外舉辦了學生基礎素養國際研究計畫（Program for International Student Assessment，簡稱PISA），主要著眼於評量學生在閱讀、數學與科學這三方面的素養。除此之外，IEA還舉辦了促進國際閱讀素養研究（Progress in International Reading Literacy Study，簡稱PIRLS），以及Teacher Education and Development Study in Mathematics（簡稱

TEDS-M）等研究，主題的種類也越來越多，而臺灣亦積極參與多項教育評比研究之中，在在顯出國際間對於教育評比的重視程度。

二、研究對象

TIMSS每四年舉辦一次大規模的國際評比調查，在進行TIMSS 1995研究的時候，當時的研究對象共分為三個群體，並涵蓋五個年級。第一個群體是以三及四年級的學童為研究對象，全球共有27個國家或地區參加。第二個群體是以七及八年級的學童為研究對象，全球共42個國家及地區參加。第三個群體是以中學最高年級學童為研究對象，對大部份的國家而言，這通常是指十二年級的學童，全球則有22個國家及地區參與此部份的評比，參與的地區則涵蓋南、北半球。雖然臺灣未能及時參加TIMSS 1995的研究，但卻能加入進行TIMSS 1999的研究，是次評比主要目的之一，是要追蹤1995年時參與TIMSS研究四年學童的那一個世代，希望能藉此提供趨勢的資料，因此研究對象是以八年級的學童為主，全球共38個國家或地區參加。1999年TIMSS的評比對象技術性的定義是從擁有最多9歲學童的兩個鄰近年級中，挑選比較高的年級“the upper of the two adjacent grades with the most 9-year-olds”為評比對象，這對一般國家而言，通常是指四年級的學童。另一組研究對象群體亦相仿，是從擁有最多9歲學童的兩個鄰近年級中，挑選比較高的年級“the upper of the two adjacent grades with the most 13-year-olds”為評比對象，這對一般國家而言是指八年級。爾後臺灣還陸續參與了TIMSS 2003和TIMSS 2007的研究。因比對於那些有參加TIMSS 1995、1999、2003及2007研究的國家或地區，研究者除了可以進行跨國比較外，尚可進行四個時間點橫跨十二年之趨勢分析。TIMSS 2003四年級的研究有25個國家或地區參加，八年級則有46個國家或地區參與。至於TIMSS 2007四年級有36個國家或地區參加，八年級有61個國家或地區參加。

三、評量的科目

TIMSS是課程導向的研究，評比的科目只聚焦於數學及科學兩個學科，為了要完整評量出學生在這兩個學科的表現，且考慮到公平性，TIMSS每一個研究都會特別設計一套評量架構作為跨國評比的依據，茲以TIMSS 2003為例介紹評比的內容。在數學科方面，評量架構是由兩個向度組織而成，第一個向度是評量的領域內容，其中包括數、代數、測量、幾何和資料分析等五個領域，而四年級內容亦相仿，但卻以規律、方程式與關係取代代數。評量架構中詳細釐定所要評量的主題內容。第二個向度是數學的認知領域，其中包括知道事實與程序、概念運用、解傳統題以及推理，每一個認知向度所涵蓋的項目尚附有詳細說明。至於科學方面的評量架構，八年級評量的領域可細分為生命科學、化學、物理、地球科學及環境科學，而四年級所評量的領域內容包括生命科學、物質科學及地球科學等領域。認知領域方面則包括事實知識、概念瞭解以及推理和分析等三個向度。該評量架構還明確標示在每一向度內每一個領域評比測驗時間分配所占的百分比（參表1），整份評量架構並會經過各參與國代表的討論及認可後公布，以達到公平的目的。關於各領域的詳細內容，請參閱Mullis等人（2003）及Martin（2005）的報告。

表1 TIMSS 2003八年級數學科各領域預期測驗時間分配的百分比

數學知識內容	
數	30 %
代數	25 %
測量	15 %
幾何	15 %
資料分析	15 %
認知能力	
知道事實與程序	15 %
概念運用	20 %
解傳統題	40 %
推理	25 %

至於TIMSS 2007研究的評量架構，基本上其內容主要是對2003年的領域做一些整合，並且對各領域的題目在測驗時間分配方面也做了一些調整，例如2007年對統計相關主題所占的百分比略微提升，以反映學術界對統計素養漸趨重視的趨勢。

四、課程架構

由於TIMSS是課程導向的評比研究，因此除了評量學生在數學與科學的知識外，還需要蒐集所有與課程相關的資訊，藉此協助解釋學生的表現，進而做跨國的比較。TIMSS研究將課程架構（curriculum framework）分為三個部份，包括計畫的課程（intended curriculum）、實施的課程（implemented curriculum）與達成的課程（attained curriculum）。計畫的課程之意涵為各國經由教育專家所訂立之法定課程標準內容，例如台灣的國民中小學九年一貫課程綱要。實施的課程是從計畫的課程出發，透過教科書作者的詮釋，以及教師以個人的瞭解及風格在課堂傳達，但實施出來的內容未必會與計畫的課程完全一致。達成的課程意指學生在課堂中參與教師所安排的各種活動，其後所能接受及吸收到的知識內容（參圖1）。三個課程之間隱含著上層影響下層的意涵。

計畫的課程之內容，需要透過設計課程問卷，委請各國代表填寫，調查並整理出各國課程所涵蓋的範圍，一方面可控制評比不致納入超出大部份國家課程範圍的題目，另一方面可藉此瞭解學童的表現是否與課程涵蓋程度有關。至於實施的課程之相關資訊，

則需要由有教導參與TIMSS研究之學童的數學和科學教師，仔細填寫教師問卷，其內容包括教師的準備、經驗、教學內容、教學方式、學校的組織及教室的資源等。科學問卷的情況較為複雜，因為部分國家的課程理念是視科學為單一整合科目，而其它國家則以分科的方式處理，上課時區分為生物、物理、化學或地球科學等科目進行教學，因此問卷又細分為整合與分科兩類教師問卷。此外，尚有一份學校問卷，由參與研究之學生所就讀學校的校長或教務主任填寫，其中詢問學校的組織及資源，以及學生在校內的態度等資料。關於達成的課程方面，一方面是以成就評量的方式評量各參與國學童的表現，題目由IEA所聘請的學科專家撰寫，並經由各國代表認同，以及題目之預試參數良好等機制，確立其信、效度後，方能進入題庫。另一方面，參與研究的學童必填寫學生問卷，調查包括社經相關背景、作業、補習以及在校內是否感到安全等情形。

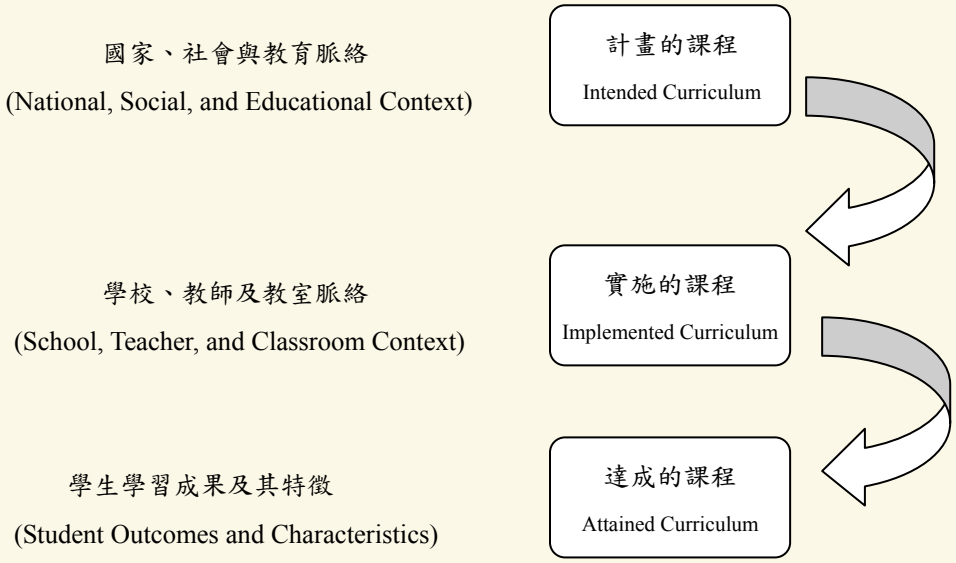


圖1　TIMSS評比研究的課程架構

過去TIMSS評量的題目以選擇題為主，同時加入了一些作答題（constructed response item），在TIMSS 2007的研究中，選擇題與作答題題數的比例已經相當接近。

五、題本設計

倘若可行，安排每位受測的學生回答所有題目，在作業上最為方便，但因TIMSS的題目太多，以2003年8年級為例，數學領域的題目有194題，科學領域有189題，總計383題，估計約需7個小時才能完成。若將數學及科學的題目分別分派到十二個題本，並且讓每位學童只回答一個題本，則平均每位學生只需要花90分鐘即可完成。明確一點來說，2003年TIMSS的題本設計（booklet design）是將數學及科學的題目分配到14個區塊（block），再將區塊分配到12個題本。

六、評分的處理

由於TIMSS評比研究需要進行跨年趨勢

的比較，若使用過去相同的題目，部份題目有可能已經過時。且因為TIMSS必須公佈部分題目作示範之用，以協助研究者詮釋學生的成就，導致部份題目曝光，故不能反覆使用所有題目，因此每一次TIMSS評比結束後，便邀請專家發展新題目取代已公開的題目，但為了進行跨年比較的緣故，仍得保留部分題目作為趨勢題（trend items），或稱之為共同題，因而產生不同年代學生回答不同題目的現象。所有題目會經過一個嚴格的翻譯過程，並將在各國進行預試，將題目去蕪存菁。因此研究者在進行TIMSS資料分析時應注意，由於學生填答之題本並不相同，其題目的難度亦不相同，所以不能直接用答對題數或答對率作為學生的成績。

因為TIMSS 2003有十二個題本14個區塊，而且要做趨勢分析，所以必需透過題目反應理論作等化的處理，將不同的題本連結在一起，才能進行跨題本及跨年的比較。TIMSS先利用共同題將各試題作校準（calibration），並採用三種IRT機率模式來估計參數，包括選擇題使用三參數模式來估計題目的難度、鑑別度及猜測的參數，二分制的短答題利用二參數模式估題目的難度及鑑別度，而多分題（得分可能為0、1、2分）則採用generalized partial credit model來進行。

在評估個人能力的時候，因每位學童只回答一份題本，其它題本的題目皆沒有作答，容易有測量誤差，因此若以單一值代表學生的能力較不恰當， 因而TIMSS採用可能分數（plausible values）的技術來進行能力估計，該技術除了依據個人試題回答的情況外，尚考慮個人背景的資料（例如：性別、看TV、功課、父母教育、家中書本數等）來推估每位學生在TIMSS整體的能力分布，然後再從該能力分佈隨機抽取五個數值作為能力的代表值，並稱之為可能分數，而且每一個可能分數都可以代表作答者的能力。該技術認為所抽取的五個可能分數的平均值可以代表該生能力的不偏估計，其分散程度可以估計出作答者能力的測量誤差。可能分數的技術源自Mislevy, Sheehan, Beaton and Johnson 等人延伸Rubin關於估計遺漏值的研究，起初是為了處理1983至1984年美國的National Assessment of Educational Progress（NAEP）調查研究而發展，其技術比較複雜，詳細的報導可參閱Mislevy（1991）及Mislevy, Beaton, Kaplan, & Sheehan（1992）等論文。該技術其後陸續應用於NAEP、TIMSS和PISA等研究中。再者，為了使一般大眾容易瞭解TIMSS的記分方式，所以又將估計出來的能力值轉換成量尺分數做報導，量尺分數的平均分數訂為500分，標準差為100分。以下表二是TIMSS 2003年分別按照數學與科學平分量尺分數排名的前五名國家。

表2 TIMSS 2003年評比數學與科學前五名國家及平分量尺分數

名次	四年級				八年級			
	數學		科學		數學		科學	
1	新加坡	594（5.6）	新加坡	565（5.5）	新加坡	605（3.6）	新加坡	578（4.3）
2	香港	575（3.2）	臺灣	551（1.7）	南韓	589（2.2）	臺灣	571（3.5）
3	日本	565（1.6）	日本	543（1.5）	香港	586（3.3）	南韓	558（1.6）
4	臺灣	564（1.8）	香港	542（3.1）	臺灣	585（4.6）	香港	556（3.0）
5	比利時（荷蘭語）	551（1.8）	英國	540（3.6）	日本	570（2.1）	愛沙尼亞	552（2.5）

註：括號內數據為標準誤。

從上表觀之，臺灣學童的表現相當不俗。不過值得注意的是，TIMSS研究以參與國家或地區學童的平均量尺分數作為名次排列的依據，可以各國學童的成績其實是一個分布，因此除了平均值之外，最好只視為參考依據之一，例如還可參考下述量尺詮釋部分的說明。

七、抽樣設計

抽樣方面，TIMSS 的目的是要將結果推廣到各國四年級及八年級的整體學童，其所採用的抽樣方法並非簡單隨機抽樣（simple random sampling），而是採用兩階段分層抽樣（2-stage stratified cluster sampling），其中第一階段採取學校抽樣，按照學校所處區域的層次抽出學校名單，學校被抽中的機率與該校該年級（四年級或八年級）學生人數有關，若抽中卻不願意參加或不符合參與資格的學校，則需要替換其它學校參與研究，為避免抽樣代表性不足的問題，被抽中的學校至少應有一定的比例參加。選定學校後，第二階段為班級抽樣，從被抽中的學校中，經扣除資源班後再隨機抽取一班，該班學生全數參與TIMSS評比研究。為了抽樣的精確度考量，各年級正式抽樣人數（sample size），原則上應包含150個學校及5000位學生。對於抽樣上達不到重重要求的國家，在國際報告中會被分開處理。

如果要將TIMSS的資料回推到母群體，則需要採用抽樣權重（sampling weights），以兼顧所使用的抽樣方式，原因在於各抽樣單位（學生）並非在相同的情況下被抽中，所以每位學生要作適當的加權，才足以代表母群體。TIMSS 有多種不同的抽樣權重係數，例如：TOTWGT、SENWGT、HOUWGT等，端視研究問題的需要來選擇使用，跨國比較常使用的方法為TOTWGT。基本上，權重與抽樣單位被抽中的機率有關，經過加權的處理可以獲得無偏誤的母群體估計值。

進行推論性統計分析時，誤差的來源一般可分為測量誤差（measurement error）及抽樣誤差（sampling error），TIMSS研究的測量誤差可透過可能分數的處理而得；抽樣誤差是抽樣時必然存在的問題，TIMSS研究是採用jackknife repeated replication技巧來進行處理。而最終的標準誤，它是上述兩種誤差的函數，是綜合測量誤差及抽樣誤差而得，詳情請參考Martin（2005）及相關技術報告。

八、量尺詮釋

TIMSS研究對於量尺意義的詮釋方面有一個特色，相當值得參考。一般評量的結果只會報導學生所得的分數，這或可作為比較之用，但對於如何改善教學以及學生的學習幫助並不大，事實上，單純報導成績對於教育的功能十分有限。除了學生所得的可能分數之外，過去在TIMSS 1999的研究中，曾採用第25、第50、第75和第90百分位數作為國際基準（international benchmark），以協助比較各國學童的成績，並扼要報導這些程度學生的表現，但由於過於簡略，這所能提供的教育資訊仍然有限。在TIMSS 2003的數學與科學的國際報告中，特別加入了一個章節報導國際基準，並詳述達到這些指標的學生能夠掌握什麼知識，其意涵在於協助教師們瞭解學生成績與能力之間的關係。

首先，TIMSS經由專家建議將學生能力以四個分數作區分（參表3），其中達625分以上者，表示該學生能力達到進階國際基準（advanced international benchmark）的程度，550分以上的學生則達到高國際基準（high international benchmark），475分以上的學生能力為中等國際基準（intermediate international benchmark），400分以上的學生能力為低國際基準（low international benchmark）。

表3 TIMSS 2003所採用的國際基準

量尺分數	國際基準
625	進階國際基準
550	高國際基準
475	中等國際基準
400	低國際基準

接著再找出那些學生的得分在這些基準的範圍內，並探討這些學生大部份掌握到什麼題目，據此撰寫達到各基準的學生之整體表現，教育界可根據學生得分在那一個基準範圍而瞭解到他們的能力，進而可考慮給予適當的教導方式，發揮出評比其實應該要有的功能。以下舉TIMSS 2003八年級數學科進階國際基準Advanced Benchmark的例子報導如下：

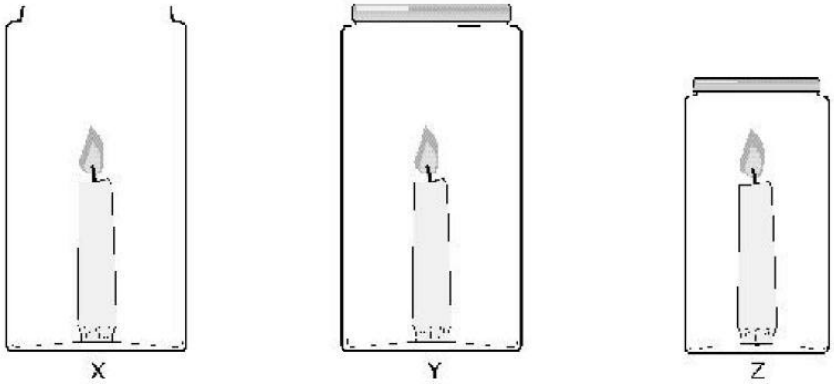
> 達到此能力的學生能夠組織資訊，做一般化的推論，解決非傳統的題目，能從給定的資料做結論並說支持的理由。他們能計算百分比的改變，並應用他們關於數字和代數的觀念與關係等知識來解決問題。
>
> 此外，這個層次的學生能夠解聯立線性方程，對簡易的情境用代數建立模式。他們能夠應用測量和幾何知識於複雜的問題情境之上，他們能從各式各樣的表格與圖表中詮釋資料，包括運用插補法與外推法。
>
> 學生能結合幾何圖形的知識來解決問題，其包含了一個以上的步驟。此知識包括了全等三角形、三角形的三角總和、內角與外角、角平分、正六邊形。他們能辨別出等半徑的弧會產生等腰三角形，能從一條線上已知的兩個座標選出在同一平面、同一直線上的其他座標，也能夠運用畢氏定理關係證明一個三角形為直角三角形。
>
> 學生能夠從數據預測其結果並運用他們對機率的理解，能夠畫出一個輪盤可以對應到一個給定表格中的數據，他們能從各式各樣的表格與圖表中解釋資料，包括插補法與外推法。他們能從已知的時間表推知出相關資訊，並為一趟特定的旅程完成一份表格，且確認此表格符合所提供的條件，他們也能根據數據求出並證明結論。（取自Mullis, Martin, Gonzalez, & Chrostowski, 2004）

九、討論

TIMSS的資料可以提供多方探討的可能性，對於瞭解各國學生在學科上的表現，以及探索學生表現差異的原因很有幫助。例如比較亞洲五個參與TIMSS 2003研究的國家或地區，其八年級生在數學科的表現依國際標

準分類後，可發現學生量尺分數在625分以上者以新加坡的學生人數最多，占44%，臺灣次之（38%），之後依序為韓國（35%）、香港(31%）及日本（24%）；成績在550分以上者，仍以新加坡為最多（77%），其次香港（73%）、韓國（70%）、臺灣（66%）及日本（62%），臺灣的排序略降；而分數低於400分的學生，在其它四個國家僅有1-2%，但是在臺灣卻有4%，顯示臺灣後段的學生相對較多，其原因為何，實值得教育界深究。另外，在四年級學生的表現方面，臺灣學生之量尺分數在625分以上者僅有16%，低於新加坡、香港及日本。此結果顯示，臺灣四年級的學生，其能力分布較為集中，但高能力者少於其它高表現的亞洲地區。這是否反映臺灣在國小階段應該留意優秀人才的培育，而國中階段多留意能力較低者，宜探究原因，革新教育政策。

因篇幅所限，茲再舉兩道題目概要討論TIMSS資料可能分映的問題。在TIMSS 2003八年級的科學作答題中，有一道公開題與蠟燭燃燒有關，題目如下：



三支完全相同的蠟燭分別置入上圖所示 X、Y 和 Z 三個罐子裡，同時點燃後，Y 和 Z 蓋上蓋子，而 X 保持開口。
哪一支蠟燭的燭火將會最先熄滅(X、Y 或 Z)？
解釋你的回答。

該題是TIMSS 2003年八年級科學方面的作答題，評量的是基本的科學推理，臺灣的國際排名是第14名，達滿分者占60%，但第一名的荷蘭得滿分者達82%，彼此有一段距離，該題的學理似乎在國二學生科學推理的能力之內，學生在這類型題目的表現，值得科學教育工作者探討其背後的原因。

另一道公開題是木星題，題目如下：

木星的體積比月球大，但是從地球上來看，則木星比月球小。為什麼？

該題臺灣的國際排名是第23名，達滿分者占66%，只比國際平均高出2%，而該題第一名的荷蘭得滿分者達88%，彼此差距達

22%。該題的原理並不困難，但卻需要用文字解釋，這是否顯示出臺灣國中二年級學生不習慣用文字解釋個人的想法，十分值得科學教育工作者探討其背後的原因。

至於TIMSS的資料還可以進行那些方面的分析，相關文獻不勝枚舉，有興趣的讀者可參考羅珮華（2003）、Chang（2006）、Shen & Tam（2009），以及Educational Research and Evaluation期刊第14卷第1期的專輯。

十、結語

進行國際教育評比的一項重要考量是公平性，例如教育評比的範圍必須為各國所接受，各種測驗評比工具必需對各國而言是具公平性的。在抽樣的部分，各國參與的學生與學校必需有充份的代表性。另外，題本設計與評分的考量，也是國際評比研究公平性不可被忽略的一環。除此之外，國際評比基本上是一個大型的標準化成就評量，因此各國在進行施測時都需要遵守劃一的步驟，測驗過程及步驟必須標準化，而且各國還需要建立監察的機制，儘可能確保該國的監考人員會嚴格依照國際的要求執行施測的程序。必須一提的是，由於各國的語言並不相同，即使是以英語為主的國家，在語言使用上亦會有文化差異，因此評比研究在翻譯上必須有嚴格的要求。TIMSS研究基本上有考量到上述各種公平性的要項，2003年的研究在進行評比前曾徵求各國提供相關科學題目，使題目來源方面更具公平性。

但由於TIMSS研究調查的範圍十分廣泛，舉凡各參與國學童在數學與科學方面的成績，以及能夠解釋他們學業表現的個人、家庭、學校與教師教學等因素，皆成為該研究調查的範疇。對於內容如此豐富的資料庫，自然會吸引許多研究者的留意，近年來，有跡象顯示有愈來愈多的研究學者進行這方面的資料分析。然而，面對一個有上千變數的資料庫，如果不瞭解其資料的性質與結構，分析時很容易會有所偏誤。研究者通常可以選擇一些變數進行資料分析，探索它們之間的關係，由於變數眾多，很容易就能找到一些關係，進而詮釋出一些結論。可是這些關係是否真的存在，抑或是偶然獲得，又或者是有其他干擾因素的存在，則非常需要研究者審慎的考量。

參考文獻

羅珮華（2003）。從「第三次國際科學與數學教育成就研究後續調查（TIMSS 1999）」結果探討國中學生學習成就與學生特質的關係：七個國家之比較。國立臺灣師範大學科學教育研究所博士論文。

譚克平（2006）。TIMSS 2003學校問卷調查的分析。載於張秋男（主編）。國際數學與科學教育成就趨勢調查2003（pp. 165-191）。台北市：國立臺灣師範大學科學教育中心。

Chang, C. N. (2006) Report of Taiwan TIMSS 2003- based on the Trends in International Mathematics and Science Study 2003. Taipei: National Taiwan Normal University.

Martin, M.O. (2005). TIMSS 2003 user guide for the international database. Chestnut Hill, MA: TIMSS & PIRLS International Study Center, Boston College.

Martin, M.O., Mullis, I.V.S., Gonzalez, E.J., & Chrostowski, S.J. (2004). TIMSS 2003 international science 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.

Mislevy, R. J. (1991). Randomization-based inference about latent variables from complex samples. Psychometrika, 56, 177-196.

Mislevy, R. J., Beaton, A.E., Kaplan, B., & Sheehan, K. M. (1992). Estimating population characteristics from sparse matrix samples of item responses. Journal of Educational Measurement, 29, 133-161.

Mullis, I.VS., Martin, M. O., Smith, T.A., Garden, R.A., Gregory, K.A., Gonzalez, E.J., Chrostowski, S.J., & O'Connor, K.M. (2003). TIMSS assessment frameworks and specifications 2003. 2nd Edition. Boston, MA: Boston College, The International Study Center.

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.

Shen, C., & Tam, H. P. (2009). The paradoxical relationship between student achievement and self-perception: A cross-national analysis based on three waves of TIMSS data. Educational Research and Evaluation: An International Journal on Theory and Practice, 14(1), 87-100.